

Repetition Makes Perfect: Recurrent Graph Neural Networks Match Message-Passing Limit

Eran Rosenbluth, Martin Grohe

rosenbluth@informatik.rwth-aachen.de, grohe@informatik.rwth-aachen.de

RWTH Aachen University

Abstract

We precisely characterize the expressivity of computable Recurrent Graph Neural Networks (recurrent GNNs). We prove that recurrent GNNs with finite-precision parameters, sum aggregation, and ReLU activation, can compute any graph algorithm that respects the natural message-passing invariance induced by the Color Refinement (or Weisfeiler-Leman) algorithm. While it is well known that the expressive power of GNNs is limited by this invariance [Morris et al., AAAI 2019; Xu et al., ICLR 2019], we establish that recurrent GNNs can actually match this limit. This is in contrast to non-recurrent GNNs, which have the power of Weisfeiler-Leman only in a very weak, "non-uniform", sense where each graph size requires a different GNN to compute with. Our construction introduces only a polynomial overhead in both time and space. Furthermore, we show that by incorporating random initialization, for connected graphs recurrent GNNs can express all graph algorithms. In particular, any polynomial-time graph algorithm can be emulated on connected graphs in polynomial time by a recurrent GNN with random initialization.

1 Introduction

Graph Neural Networks Message-Passing Graph Neural Networks (GNNs) (Kipf and Welling 2017; Gilmer et al. 2017) is a class of graph-processing architectures commonly used in tasks of learning on graphs. As such, characterizing their expressivity is of great importance.

A GNN is a finite sequence of operations, called *layers*, applied in parallel and in sync to each node. Its inputs are graphs whose nodes are assigned an initial feature-vector of certain dimension. In the broadest theoretical setting the vector is over the real numbers, however, most expressivity studies consider more restrictive domains: Some consider integers, some consider a compact segment $[a, b] \subset \mathbb{R}$, and most consider a finite domain. A layer starts with constructing a *message* from each neighbor, often being simply the neighbor's current value, although more sophisticated algorithms can also be used. Importantly, messages bear no identification of the sending node. An aggregation algorithm, typically dimension-wise sum; avg; or max, is then applied to the multiset of messages, producing a fixed-dimension value. Finally, a combination algorithm in the form of a *Multilayer Perceptron* (MLP) is applied to the aggregation value and the node's current value, producing the node's new

value. For node-level tasks, a node's value after the application of the last layer is considered the output of the GNN for that node. For graph embeddings, the nodes' final values are aggregated, and an MLP is applied to the aggregation value, producing the GNN's output for the graph.

The node-centric, node-indifferent, nature of the algorithm means every GNN can technically be applied to graphs of all sizes and GNNs are isomorphism-invariant. The MLP part of the layers gives GNNs their learnability qualities.

It is common for GNNs to use the same aggregation type in all layers. Those that use sum are denoted *Sum-GNNs*.

The expressivity boundaries of GNNs can be attributed to the combination of three factors:

1. The message-passing scheme. A main consequence of this is a distinguishing power that cannot exceed that of the Color Refinement algorithm - or 1-dimensional *Weisfeiler-Leman* (1-WL) as it is sometimes referred to (Xu et al. 2019; Morris et al. 2019; Aamand et al. 2022).
2. The fixed number of layers-executions. An obvious effect of that is a fixed information-radius for each node's computation. Another effect is impossibility to enhance the expressivity of the layers' algorithms (see next) by means of repetition.
3. The combination and aggregation functions that make GNNs' layers. For reasons of learnability and runtime performance, these are of specific classes as mentioned above. While MLPs are universal approximators of continuous functions on compact domains (Hornik, Stinchcombe, and White 1989), their expressivity is very limited for non-compact domains e.g. no MLP can approximate $\forall x \in \mathbb{N} \ x \mapsto x^2$. As for the aggregation functions, all mentioned choices potentially lose information about the multiset of neighbors' messages.

Expressivity Notions In more than a few studies (Chen et al. 2019; Grohe 2023), an architecture is considered to be expressive of a function f if and only if it *non-uniformly* expresses f : For every graph size n there exists a model M_n of the architecture that approximates f on all graphs **of that size**. Non-uniform expressivity is not only a weak guarantee in theory, but it also has limited relevance to practice: First, non-uniform expressive GNNs that are non-polynomial in size are too large, before anything else. Second, all non-uniform expressive GNNs may succeed at inference time

only when the input-graph size does not exceed the sizes of graphs in training time, otherwise they fail miserably - as implied by the theory and observed in experiments (Rosenbluth, Toenshoff, and Grohe 2023; Rosenbluth et al. 2024).

The strongest and most meaningful notion of expressivity, both to theory and practice, and the one that we use in this paper, is *uniform* expressivity. A GNN architecture is said to uniformly express a function f if and only if there exists (at least) one specific model of that architecture that approximates f **on graphs of all sizes**. It implies that the expressing GNN computes a function that generalizes in size, rather than computing a sort of lookup table for the finitely many graphs of a specific size. From a practical standpoint, uniform expressivity means it may be possible to train a GNN model on graphs of smaller sizes and this model will approximate the target function also on graphs of larger sizes. We are not aware of any meaningful tight bounds shown for the uniform expressivity of GNNs. Obviously, GNNs cannot uniformly express functions that depend on an unbounded information-radius, but their limitations go beyond that: Previous works have shown that common GNN architectures cannot uniformly express basic regression and classification functions even when they are expressible by another GNN architecture and the required information-radius is only 2 (Rosenbluth, Toenshoff, and Grohe 2023; Grohe and Rosenbluth 2024).

Recurrent Graph Neural Networks Recurrent GNNs (Scarselli et al. 2008; Gallicchio and Micheli 2010) are similar to GNNs in that a recurrent GNN consists of a finite sequence of layers and those comprise the same algorithms of (non-recurrent) GNN layers. However, in a recurrent GNN the sequence can be reiterated a number of times that depends on the input. Recurrent GNNs have been used successfully in practice (Li et al. 2016; Selsam et al. 2016; Bresson and Laurent 2018; Tönshoff et al. 2023) and are considered promising architectures for solving various learning tasks. In terms of their uniform expressivity, recurrent GNNs are still limited by factor (1) as they still consist of local algorithms; by definition factor (2) does not apply to them; and the question is to what extent recurrence can mitigate factor (3). Recurrent MLPs with ReLU activation are proved to be Turing-complete (Siegelmann and Sontag 1992), implying that a layer’s expressivity may be increased by reiterating it. However, it has not been clear thus far if reiteration can recover the information lost in aggregation.

A recurrent message-passing aggregate-combine architecture, where the nodes are aware of the graph size, was shown to be as expressive as message-passing can be (Pfluger, Cucala, and Kostylev 2024). However, the layers’ functions there are not restricted to be computable, let alone an MLP and a common aggregation function, making the result only an upper bound for computable recurrent GNNs. For recurrent sum-aggregation GNNs, it has been proven that a single layer GNN can distinguish any two graphs distinguishable by Color Refinement (Bravo, Kozachinskiy, and Rojas 2024). However, the proof is existential, the activation function of the GNN must not be ReLU, and most importantly, the GNN has a parameter whose value must

be a real number - of infinite precision - making it incomputable. A tight logical bound for a computable recurrent GNN architecture, named by the authors an ‘R-Simple AC-GNN[F]’, is proven in (Ahvonnen et al. 2024). The architecture is defined to operate only with fixed-length float values, making it limited in one aspect, and to aggregate multisets of such values always in order (e.g. ascending), making it sophisticated in another. All in all, expressivity-wise it is a strictly weaker architecture, and essentially different, than the recurrent GNNs we study. Moreover, it is characterized in different terms than the ones we use to tightly characterize our architectures.

Message-Passing Limit By their definition, the uniform expressivity of computable recurrent GNNs is upper-bounded by the expressivity of a general message-passing algorithm with id-less nodes that are aware of the graph-size: A single recurring algorithm operating in parallel at each node, whose input at each recurrence is the node’s value and a multiset of messages from the node’s neighbors, and its output is a new value and a message to send. Importantly, the same message is broadcasted to all neighbors, and messages are received with no identification of the sending node. Recurrent GNNs are then a specific case, where the recurring algorithm is an MLP composed on an aggregation of the messages, rather than a general algorithm operating on the multiset of messages.

A well-studied representation of a node defines another important upper-bound: The class of all algorithms $\mathcal{A}(G, v)$, for a graph and node, that are invariant to the Color Refinement (CR) (or 1-Weisfeiler-leman) representation of the node. The Color Refinement procedure goes back to (Morgan 1965; Weisfeiler and Leman 1968), see also (Grohe 2021) and Section 3.

New Results We prove meaningful tight bounds for the uniform expressivity of computable recurrent GNNs - with finite precision weights and ReLU activation. All the reductions are achieved with polynomial time and space overhead. We assume a finite input-feature domain, and that the original feature is augmented with the graph-size value. There, we show that a recurrent single-layer sum-aggregation GNN can compute the following (see Figure 1 for an illustration):

1. Node-level functions:
 - a. Any algorithm that is invariant to the Color Refinement (CR) value of the node. (Thm 4.2)
 - b. Any message-passing algorithm. (Implied by (a))
 - c. When adding global aggregation, any algorithm that is invariant to the Weisfeiler-Leman (WL) value of the node. (Thm. 5.3)
 - d. For connected graphs, with random node initialization, any algorithm that is isomorphism-invariant (Cor. 5.2).
2. For connected graphs, any computable graph embedding that is invariant to the CR value of the graph. (Thm. 5.1)

Roadmap In Section 3 we describe the color of a node and its relation to the message-passing scheme, and use it to define our upper-bound function classes. In Section 4 we

Recurrent GNNs + Random Node Initialization \equiv^* Isomorphism-Invariant Node Functions

*holds only for connected graphs, probabilistic

Recurrent GNNs + Virtual Nodes \equiv WL-Invariant Node Functions

Recurrent GNNs \equiv CR-Invariant Node Functions \equiv Message-Passing Algorithms

Non-Recurrent GNNs

Figure 1: Uniform expressivity hierarchy. 'CR' and 'WL' are acronyms for Color Refinement and Weisfeiler-Leman. The results in this paper are the equivalencies. 'Node Functions' means computable functions $f(G, v)$ operating on a graph and node.

define our recurrent GNN architecture and describe the reduction from the upper-bound classes through intermediate models and down to our architecture. In Section 5 we extend our results to graph embeddings and to recurrent GNN architectures that go beyond pure message-passing.

2 Preliminaries

By $\mathbb{N}; \mathbb{Q}; \mathbb{R}$ we denote the natural; rational; and real numbers respectively. We define $\mathbb{Q}_{[0,1]} := \{q : q \in \mathbb{Q}, 0 \leq q \leq 1\}$. Let $v \in \mathbb{R}^d$ be a d -dimension vector, we denote by $v(i)$ the value at position i , and by $v[a, b] := (v_a, \dots, v_b)$ the sub-vector from position a to b . Let $v \in \mathbb{R}^d$ and $a \in \mathbb{R}$, we define $v + a := v(1) + a, \dots, v(d) + a$. Let $v_1, v_2 \in \mathbb{R}^d$ be d -dimension vectors, we define $v_1 + v_2 := v_1(1) + v_2(1), \dots, v_1(d) + v_2(d)$. For vectors $u \in \mathbb{R}^m, v \in \mathbb{R}^n$ we define $u, v := u(1), \dots, u(m), v(1), \dots, v(n)$. For a set S , we denote the set of all finite multisets with elements from S by $\binom{S}{*}$. We denote the set of all finite tuples with elements from S by S^* . For a vector $v \in \mathbb{R}^d$ we define $\dim(v) := d$, and for a matrix $W \in \mathbb{R}^{d_1 \times d_2}$ we define $\dim(W) := (d_1, d_2)$.

We define $\mathcal{B} := \{0, 1\}^*$ the set of all finite binary strings, and $\mathcal{B}_k := \{0, 1\}^k$ the set of all binary strings of length k . For $x \in \mathcal{B}_k$ we define $|x| := k$.

A (vertex) *featured graph* $G = \langle V(G), E(G), S, Z(G) \rangle$ is a 4-tuple being a graph with a *feature map* $Z(G) : V(G) \rightarrow S$, mapping each vertex to a value in some set S . Let $v \in V(G)$, we denote $Z(G)(v)$ also by $Z(G, v)$. We define the *order*, or *size*, of G , $|G| := |V(G)|$. A *rooted graph* is a pair (G, v) where G is a graph and $v \in V(G)$ is a vertex. We denote the set of graphs featured over S by \mathcal{G}_S and the set of all featured graphs by \mathcal{G}_* . Note that in this paper we focus on the graph domain \mathcal{G}_B i.e. single-dimension featured graphs where the feature is a bit-string. However, our results apply to multi-dimensional featured graphs as well: A node's tuple can be encoded into one dimension during pre-processing, or, alternatively, the construction in our proofs can be extended such that the initial stage of the computation includes encoding the tuple into one dimension.

We denote the set of all feature maps that map to some set

T by \mathcal{Z}_T , and we denote the set of all feature maps by \mathcal{Z}_* . Let $\mathcal{G} \subseteq \mathcal{G}_*$, a mapping $f : \mathcal{G} \rightarrow \mathcal{Z}_*$ to new feature maps is called a *feature transformation*.

A *message-passing algorithm*¹ is a pair $C = (A, f)$, $A \in \mathcal{B}$, $f : \mathcal{B}^2 \rightarrow \mathcal{B}^2$ comprising an initial state and a computable function. It defines a feature transformation $C : \mathcal{G}_B \rightarrow \mathcal{Z}_B$ as follows: Define

$\forall G \in \mathcal{G}_B \forall v \in V(G) C^{(0)}(G, v) := \theta(A, |G|, Z(G, v)), \emptyset$ for some encoding θ of the initial state; the graph size; and the initial feature, and \emptyset denoting the empty string. Then, define $\forall i > 0 C^{(i+1)}(G, v) :=$

$$f(C^{(i)}(G, v)(1), \mu(\{C^{(i)}(G, w)(2) : w \in N_G(v)\}))$$

for some multiset encoding μ . Define the first iteration when a 'finished' indicator turns 1:

$$I_v := \begin{cases} \min(i : C^{(i)}(G, v)(1)(1) = 1) & \text{min exists} \\ \infty & \text{otherwise} \end{cases}$$

Finally, $C(G, v) := C^{(I_v)}(G, v)(2)$, if I_v is defined.

Throughout the paper we refer to Multilayer Perceptrons (MLPs), meaning specifically ReLU-activated MLPs, formally defined in the appendix. A d dimension *recurrent MLP* F is an MLP of I/O dimensions $d; d$. It defines an iterative function $f_F^{(t)}$, $t \in \mathbb{N}$ such that $f_F^{(0)}(x) := x$, $\forall t > 0 f_F^{(t)}(x) := f_F(f_F^{(t-1)}(x))$.

3 Message-Passing Information

In this section, we give a precise technical description of the limits of message-passing algorithms using the *Color Refinement* procedure. It aims to describe the maximal "message-passing information" each node can obtain. It will be convenient to refer to this message-passing information as the "color" of a node. Note, however, that the message-passing colors are complex objects, nested tuples of multisets, which will later be compactly represented by a directed acyclic graph (dag).

¹Our definition is equivalent to any distributed algorithm where the initial input includes the graph size, and the input from a node's neighbors is a multiset - with no ids or order

Definition 3.1. Let $G \in \mathcal{G}_{\mathcal{B}_k}$. For every $t \geq 0$ and $v \in V(G)$, we define the message-passing color of v after t rounds inductively as follows: The initial color of v is just its feature in G , that is, $\text{mpc}_G^{(0)}(v) := Z(G, v)$. The color of v after $(t+1)$ rounds is the color of v after t rounds together with the multiset of colours of v 's neighbours, that is,

$$\text{mpc}_G^{(t+1)}(v) := (\text{mpc}_G^{(t)}(v), \{\{\text{mpc}_G^{(t)}(w) \mid w \in N_G(v)\}\}).$$

Moreover, we define the final color of v to be

$$\text{mpc}_G(v) := \text{mpc}_G^{2|G|}(v).$$

For all $t, n, k \in \mathbb{N}$, we let

$$\begin{aligned} \text{MPC}_{n,k}^{(t)} &:= \{\text{mpc}_G^{(t)}(v) \mid G \in \mathcal{G}_{\mathcal{B}_k}, |G| = n, v \in V(G)\}, \\ \text{MPC}^{(t)} &:= \bigcup_{n,k \in \mathbb{N}} \text{MPC}_{n,k}^{(t)}, \quad \text{MPC} := \bigcup_{t \in \mathbb{N}} \text{MPC}^{(t)}. \end{aligned}$$

While most applications of Color Refinement are mainly interested in the partition of the vertices into color classes, for us, the actual colors carrying the message-passing information are important. If written as strings in a straightforward manner, the colors will become exponentially large (up to size $\Omega(n^t)$). We may also view the colors by trees, which are still exponentially large. We introduce a polynomial-size dag representation for each color $c \in \text{MPC}_{n,k}^{(2n)}$, denoted $D(c)$. For the full description of the dag construction, please refer to the appendix.

A feature transformation $F : \mathcal{G}_{\mathcal{B}} \rightarrow \mathcal{Z}_{\mathcal{B}}$ is *message-passing-invariant* (for short: *mp-invariant*) if for all graphs $G, H \in \mathcal{G}_{\mathcal{B}}$ of the same order $|G| = |H|$ and nodes $v \in V(G), w \in V(H)$, if $\text{mpc}_G^{(t)}(v) = \text{mpc}_H^{(t)}(w)$ for all $t \geq 1$ then $F(G, v) = F(H, w)$. By induction on the number of message-passing rounds, every feature transformation computed by a message-passing algorithm is mp-invariant. The converse is implied by the fact that, clearly, $D(\text{mpc}_G(v))$ can be constructed by a message-passing algorithm, together with the following lemma which asserts that, remarkably, $D(\text{mpc}_G(v))$ suffices to compute any mp-invariant feature transformation.

Lemma 3.2. Let $F : \mathcal{G}_{\mathcal{B}} \rightarrow \mathcal{Z}_{\mathcal{B}}$ be a computable feature transformation. Then F is mp-invariant if and only if there is an algorithm that computes $F(G, v)$ from $\text{mpc}_G(v)$. More precisely, there is an algorithm that, given $D(\text{mpc}_G(v))$, computes $F(G, v)$, for all $G \in \mathcal{G}_{\mathcal{B}}$ and $v \in V(G)$. Furthermore, if F is computable in time $T(n)$ then the algorithm can be constructed to run in time $T(n) + \text{poly}(n)$, and conversely, if the algorithm runs in time $T(n)$ then F is computable in time $T(n) + \text{poly}(n)$.

The crucial step towards proving this lemma is to reconstruct a graph from the message-passing color: Given $D(\text{mpc}_G(v))$, we can compute, in polynomial time, a graph G' and a node v' such that $\text{mpc}_{G'}(v') = \text{mpc}_G(v)$. This nontrivial result is a variant of a theorem due to (Otto 1997). Once we have this reconstruction, Lemma 3.2 follows easily.

4 Main Result

We would like to characterize the expressivity of R-GNNs, which operate on rational numbers, in terms of computable

functions i.e. algorithms that operate on bit-strings. We use two encodings of the latter representation by the former: Rational Quaternary and Rational Binary, the reasons for which reside in the proof of Lemma 4.8. Let

$$\text{RQ} := \{\sum_{i=1}^k a_i 4^{-i} : \forall j \in [k] a_j \in \{1, 3\}, k \in \mathbb{N}\}$$

$$\text{RB} := \{\sum_{i=1}^k a_i 2^{-i} : \forall j \in [k] a_j \in \{0, 1\}, k \in \mathbb{N}\}$$

, then define the encoding operations

$$\text{rq} : \mathcal{B} \rightarrow \text{RQ}, \quad \text{rq}(b_1, \dots, b_k) := \sum_{i=1}^k (2b_i + 1) 4^{-i}$$

$$\text{rb} : \mathcal{B} \rightarrow \text{RB}, \quad \text{rb}(b_1, \dots, b_k) := \sum_{i=1}^k b_i 2^{-i}$$

For vectors of binary strings we may use the rq, rb notations to denote the element-wise encoding, that is, $\forall (B_1, \dots, B_l) \in \mathcal{B}^l$

$$\text{rq}(B_1, \dots, B_l) := (\text{rq}(B_1), \dots, \text{rq}(B_l)),$$

$$\text{rb}(B_1, \dots, B_l) := (\text{rb}(B_1), \dots, \text{rb}(B_l))$$

We are now ready to define the recurrent GNN architecture that is the subject of our main result. Part of the definition is the initial input provided to it. We choose to include the maximum feature length (across all nodes) k in that input, as this allows us later to construct a single GNN for all $G \in \mathcal{G}_{\mathcal{B}}$, as stated in Theorem 4.2. Alternatively, we could waive having k in the input, make it instead a parameter of the architecture, and restrict the statement in Theorem 4.2 to all $G \in \mathcal{G}_{\mathcal{B}_k}$.

Definition 4.1. A Recurrent Sum-GNN (*R-GNN*) $N = (A, F)$ of dimension d is a pair comprising a constant initial-state vector $A \in \mathbb{Q}^{d-3}$ and an MLP F of I/O dimensions $2d; d$. Dimension d is a 'computation finished' indicator, and dimension $d-1$ holds the computation result. It defines a feature transformation $N : \mathcal{G}_{\mathcal{B}} \rightarrow \mathcal{Z}_{\mathcal{B}}$ as follows: Let $G \in \mathcal{G}_{\mathcal{B}}$, let $k := \max(|b| : b \in \text{img}(Z(G)))$ the maximum length over the binary-string features of the vertices in G , and let $v \in V(G)$. Define the initial value of N to be the rational vector comprising the graph size; max feature length; initial feature; and initial state, that is,

$$N^{(0)}(G, v) := |G|, k, \text{rb}(Z(G, v)), A$$

Define the value of N after $t > 0$ iterations to be

$$N^{(t)}(G, v) := F(N^{(t-1)}(G, v), \sum_{w \in N_G(v)} N^{(t-1)}(G, w))$$

Define the first iteration when 'finished' turns 1

$$I_v := \begin{cases} \min(i : N^{(i)}(G, v)(d) = 1) & \text{min exists} \\ \infty & \text{otherwise} \end{cases}$$

Then,

$$N(G, v) := \begin{cases} \text{rb}^{-1}(N^{(I_v)}(G, v)(d-1)) & I_v \in \mathbb{N} \\ \text{undefined} & \text{otherwise} \end{cases}$$

the binary string represented in rational-binary encoding at position $(d-1)$, when the 'finished' indicator turns 1.

We define a time measure $T_N(G, v) := I_v$, and

$$T_N(G) := \max(I_v : v \in V(G))$$

We say that an R-GNN N uses time $T(n)$, for a function $T : \mathbb{N} \rightarrow \mathbb{N}$, if for all graphs G of order at most n it holds that $T_N(G) \leq T(n)$. We define $L_N(G, v)$ to be the largest bit-length over all parameters' and neurons' values of F , at any point of the computation for v , and we define

$$L_N(G) := \sum_{v \in V(G)} L_N(G, v)$$

We say that an R-GNN N uses space $S(n)$, for a function $S : \mathbb{N} \rightarrow \mathbb{N}$, if for all graphs G of order at most n it holds that $L_N(G) \leq S(n)$.

Note that reaching a fixed point is not required for our results, hence it is not part of R-GNNs termination definition. However, the R-GNN we construct in the proof of Lemma 4.8 does have that property i.e. $I_v \in \mathbb{N} \Rightarrow \forall t \geq I_v \ N^{(t)}(G, v)[d-1, d] = N^{(I_v)}(G, v)[d-1, d]$, which may be useful in practice and for relation to logic.

By Theorem B.2, having the graph size in the input is a must for maximum expressivity. For recurrent GNNs with *global readout* (see Section 5), this requirement is removed since such GNNs can compute the size.

Our main theorem refers to mp-invariant functions. However, since every message-passing algorithm is mp-invariant and since R-GNNs are specific message-passing algorithms, we have that R-GNNs are expressivity-wise equivalent also to message-passing algorithms.

Theorem 4.2. Let $F : \mathcal{G}_{\mathcal{B}} \rightarrow \mathcal{Z}_{\mathcal{B}}$ be a computable feature transformation. Then F is mp-invariant if and only if there is an R-GNN N such that

$$\forall G \in \mathcal{G}_{\mathcal{B}} \ \forall v \in V(G) \ N(G, v) = F(G)(v)$$

Furthermore, if F is computable in time $T(n)$ and space $S(n)$ then N uses time $O(T(n)) + \text{poly}(n)$ and space $O(S(n)) + \text{poly}(n)$.

Intermediate Reductions

Our proof of Theorem 4.2 reduces an mp-invariant function to an R-GNN through a sequence of three intermediate computation models, see Figure 2 for a detailed illustration. The models operate on bit-strings, hence we define bit-encodings for data entities that appear in the models' definitions.

We define $\delta : \text{MPC} \rightarrow \mathcal{B}$ to be an encoding of space complexity $O(n^3 \log n + kn)$ for all $c \in \text{MPC}_{n,k}^{(t)}$, $t \leq 2n$, such that all required operations can be done in polynomial time. Following the construction description of $D(\text{mpc}_G(v))$ in Section A, it is evident that there exists such an encoding.

We define $\mu : \binom{\mathcal{B}}{*} \rightarrow \mathcal{B}$ to be a multiset encoding such that the elements can be encoded and decoded in linear time by a Random Access Machine (RAM). We define $\theta : \mathcal{B}^* \rightarrow \mathcal{B}$ to be a tuple encoding such that the elements can be encoded and decoded in linear time by a RAM. Clearly, such encodings exist.

For $l, c, k \in \mathbb{N}$ we define $\theta_{k,c}^{(l)} : (\mathcal{B}_k)^l \rightarrow \mathcal{B}$ to be a tuple encoding of l bit-strings of length at most k , of space complexity $O(l(\log(c) + k))$, such that the elements can be encoded and decoded in linear time by a RAM, and, most importantly, the separation between the elements is

preserved under summation of c such encodings: For all $(x_1, \dots, x_c), x_i \in (\mathcal{B}_k)^l$ it holds that

$$\Sigma_{i=1}^c (\theta_{k,c}^{(l)}(x_i)) = \theta_{k,c}^{(l)} \left(\Sigma_{i=1}^c (x_i(1)), \dots, \Sigma_{i=1}^c (x_i(l)) \right)$$

A straightforward encoding that reserves $\log(c2^k)$ bits for each one of the l parts satisfies the requirements.

Definition 4.3. An MPC Graph Algorithm (MPC-GA) $C = (M)$ is simply a Turing machine. It defines a feature transformation $C : \mathcal{G}_{\mathcal{B}} \rightarrow \mathcal{Z}_{\mathcal{B}}$ as follows:

Let $G \in \mathcal{G}_{\mathcal{B}}, v \in V(G)$, then $C(G, v) := M(\delta(\text{mpc}_G(v)))$.

Let F be an mp-invariant function. By Lemma 3.2, there exists an MPC-GA that computes F with polynomial time and space overhead. Hence, the first stage in the reduction sequence in Figure 2 is already proven. The next step is to translate MPC-GA - whose input is already the color of a vertex - to a distributed algorithm that has to gather that information before applying the core algorithm to it.

Definition 4.4. Let $C = (A, f)$, $A \in \mathcal{B}$, $f : \mathcal{B}^2 \rightarrow \mathcal{B}^2$ be a pair, comprising an initial state and a computable function. It defines a feature transformation $C : \mathcal{G}_{\mathcal{B}} \rightarrow \mathcal{Z}_{\mathcal{B}}$ as follows: Let $G \in \mathcal{G}_{\mathcal{B}}$ and $v \in V(G)$, then define

$$C^{(0)}(G, v) := \theta(A, |G|, \delta(Z(G, v))), \delta(Z(G, v))$$

a 2-dimension vector, the first binary string being the tuple encoding of the initial state; graph size; and encoding of the initial feature, and the second being only an encoding of the initial feature. Define $\forall t \geq 0 \ C^{(t+1)}(G, v) :=$

$$f(C^{(t)}(G, v)(1), \mu(\{C^{(t)}(G, w)(2) : w \in N_G(v)\}))$$

the value after $t+1$ iterations. Finally, define

$$C(G, v) := C^{(2|G|+1)}(G, v)(2)$$

That is, the final output is the second output of f after the $2|G| + 1$ iteration. We say that $C = (A, f)$ is a Message Passing Limited Graph Algorithm (MP-LGA) if

$$\forall G \in \mathcal{G}_{\mathcal{B}_k} \ \forall v \in V(G) \ \forall t \in [2|G| + 1]$$

$$|\mu(\{C^{(t-1)}(G, w)(2) : w \in N_G(v)\})| \leq O(3k|G|^4)$$

, that is, the bit-length of the multiset of neighbors' messages does not exceed $3k|G|^4$.

Lemma 4.5. Let $C = (M)$ be an MPC-GA, then there exist $A \in \mathcal{B}$, $f : \mathcal{B}^2 \rightarrow \mathcal{B}^2$ such that $C' = (A, f)$ is an MP-LGA and $\forall G \in \mathcal{G}_{\mathcal{B}} \ \forall v \in V(G) \ C(G, v) = C'(G, v)$. Furthermore, C' incurs polynomial time and space overhead.

Next, we reduce the MP-LGA model to a model where the input from neighbors is **the sum** of the multiset of neighbors messages rather than the multiset itself. This addresses the first main obstacle of the overall reduction: To recover the information lost by the sum-aggregation and reconstruct the multiset of messages.

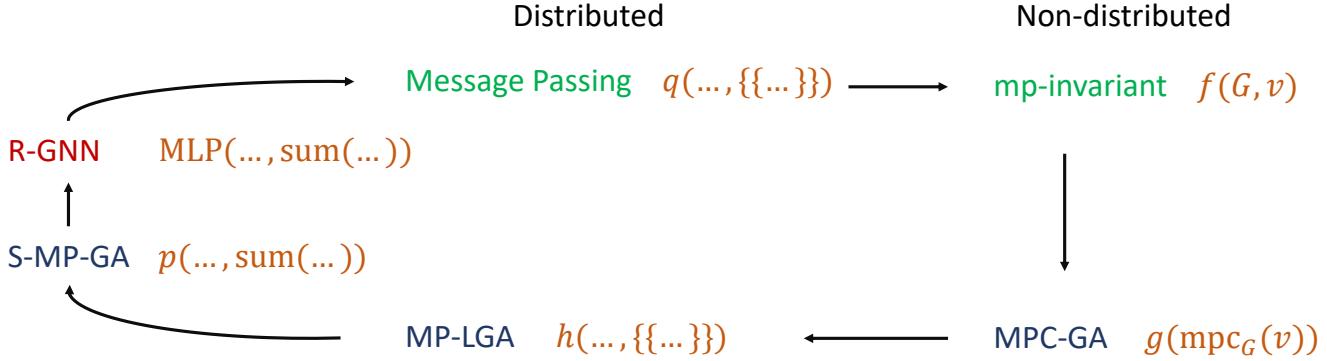


Figure 2: An overview of the reduction sequence from message-passing algorithms and mp-invariant functions to R-GNNs. Every message-passing algorithm is mp-invariant - by induction on the number of iterations. Then, starting from the mp-invariant class and moving clockwise, the reductions correspond to Lemma 3.2; Lemma 4.5; Lemma 4.7; and Lemma 4.8.

Definition 4.6. Let $C = (A, f)$, $A \in \mathcal{B}$, $f : \mathcal{B}^4 \rightarrow \mathcal{B}^4$ be a pair, comprising an initial state and a computable function. It defines a feature transformation $C : \mathcal{G}_{\mathcal{B}_k} \rightarrow \mathcal{Z}_{\mathcal{B}}$ as follows: Let $G \in \mathcal{G}_{\mathcal{B}_k}$, $v \in V(G)$, then define

$$C^{(0)}(G, v) := \theta(|G|, Z(G, v), A), \theta_{k, |G|}^{(2)}(1, Z(G, v)), 0, 0$$

a vector of 4 binary strings, the first being an encoding of the graph size; initial feature; and initial state, the second being an encoding of 1 and the initial feature, and the 3rd and 4th representing the final result and a 'finished' indicator. Define $\forall t \geq 0 \ C^{(t+1)}(G, v) :=$

$$f(C^{(t)}(G, v)(1), \sum_{w \in N(v)} C^{(t)}(G, w)(2), C^{(t)}(G, v)[3, 4])$$

the value after $t + 1$ iterations. Define

$$I_v := \begin{cases} \min(i : C^{(i)}(G, v)(4) = 1) & \text{min exists} \\ \infty & \text{otherwise} \end{cases}$$

$$C(G, v) := \begin{cases} C^{(I_v)}(G, v)(3) & I_v \in \mathbb{N} \\ \text{undefined} & \text{otherwise} \end{cases}$$

That is, the result is the binary string at position 3, when the 'finished' indicator turns 1. We say that $C = (A, f)$ is a Sum MP Graph Algorithm (S-MP-GA) if $\forall G \in \mathcal{G}_{\mathcal{B}_k} \ \forall v \in V(G) \ \forall t \in [2|G| + 1]$ it holds that $|\sum_{w \in N(v)} C^{(t-1)}(G, w)(2)| \leq O(3k|G|^4)$ i.e. the bit-length of the sum of neighbors' messages is bounded by $3k|G|^4$.

Besides having a sum aggregation, an S-MP-GA differs from an MP-LGA in two technical properties:

1. A message sent by a vertex consist of two parts, 'count' and 'value', rather than one. This is in order to count the number of sending vertices, which is useful later.
2. Two dimensions are used solely to define the final value, to make S-MP-GAs similar in that regard to R-GNNs.

The next lemma is a key step in our sequence of reductions.

Lemma 4.7. Let $C = (A, f)$ be an MP-LGA then there exist A', f' such that $C' = (A', f')$ is an S-MP-GA and $\forall G \in \mathcal{G}_{\mathcal{B}} \ \forall v \in V(G) \ C'(G, v) = C(G, v)$. Furthermore, C' incurs polynomial time and space overhead.

Reduction to R-GNN

Finally, we reduce S-MP-GAs to R-GNNs. The essential difference between the models is the recurring algorithm: In an S-MP-GA it can be any computable function i.e. a Turing machine, while in an R-GNN it is restricted to be an MLP.

Lemma 4.8. Let $C = (B, h)$ be an S-MP-GA, then there exists an R-GNN $N = (A, F)$ such that $\forall G \in \mathcal{G}_{\mathcal{B}} \ \forall v \in V(G) \ C(G, v) = N(G, v)$. Furthermore, N incurs polynomial time and space overhead.

Note that an R-GNN is, in a way, an extension of a recurrent MLP to the sum-aggregation message-passing setting. In (Siegelmann and Sontag 1992) it is shown that recurrent MLPs are Turing-complete. Let M be a Turing-machine that computes h , we would like to use the result in (Siegelmann and Sontag 1992) and emulate M using the recurrent MLP in an R-GNN. Yet, this requires overcoming two significant gaps:

1. An encoding gap. In (Siegelmann and Sontag 1992) the emulation of a Turing machine is done by emulating a two-stack machine where a stack's content is always represented as a value in RQ. Since RQ is not closed under summation, a naive attempt to use the sum of the neighbors' stacks directly - as input to the Turing machine emulation - is doomed to fail: The sum may be an invalid input and consequently the output will be wrong. To overcome this, we translate outgoing messages from RQ to RB - which is closed under summation, and we translate the incoming sum-of-messages back to RQ. The translations are implemented by two dedicated recurrent sub-MLPs of F .
2. A synchronization gap. In an S-MP-GA computation, nodes are synchronized by definition: The i^{th} recurrence's input is the the sum of results of the $i - 1$ application of h to each of the neighbors. However, in an emulating R-GNN that is based on (Siegelmann and Sontag 1992), every recurrence corresponds merely to a Turing machine step. As different nodes may require a different number of steps to complete the computation of a single

application of h , when a node finishes its i^{th} emulation of h , in the t^{th} recurrence, its external input in recurrence $t + 1$ is not necessarily the sum of its neighbors' i^{th} result, and it is unknown in what recurrence it will be. To overcome this, we augment the recurrent MLP described thus far with a recurrent sub-MLP of F that implements a synchronization algorithm across all nodes. That sub-MLP runs for the same number of recurrences for all nodes, hence, in itself it is synchronized.

Overall, the recurrent MLP F consists of 8 recurrent sub-networks, each fulfilling a different task. See Figure 5 in the appendix for an outline of F 's structure. At each iteration only one sub-network, other than the synchronizer, executes its task while the others execute a computation that does not affect their main outputs. See the appendix for further details.y

5 Further Results

Our main theorem, Theorem 4.2, has a number of interesting variants and implications. First, it has a version for *graph embeddings* $F : \mathcal{G}_B \rightarrow \mathcal{B}$. A graph-embedding R-GNN $N = (R, M)$ comprises an R-GNN R and an MLP M . It applies R , averages the nodes' values, and applies M . More formally, let I_v be the last iteration of R on a node v , then

$$\forall G \in \mathcal{G}_B \quad N(G) := \text{rb}^{-1} \left(M \left(\frac{1}{|G|} \sum_{v \in V(G)} R^{(I_v)}(G, v) \right) \right)$$

We say that two graphs are *mp-indistinguishable*, or indistinguishable by Color Refinement, if the same message-passing colors appear with the same multiplicities, that is, $\forall t \geq 0$

$$\{\{\text{mpc}_G^{(t)}(v) \mid v \in V(G)\}\} = \{\{\text{mpc}_H^{(t)}(v) \mid v \in V(H)\}\}$$

Note that this implies $|G| = |H|$. A function $F : \mathcal{G}_B \rightarrow \mathcal{B}$ is *mp-invariant* if for all mp-indistinguishable graphs G, H we have $F(G) = F(H)$. Clearly, all graph embeddings computable by R-GNNs are mp-invariant. The following states the converse, which holds only for connected graphs, a limitation proved in Theorem C.1.

Theorem 5.1. *Let $\mathcal{CG}_B \subset \mathcal{G}_B$ be the subset of connected graphs in \mathcal{G}_B , and let $F : \mathcal{CG}_B \rightarrow \mathcal{B}$ be computable. Then, F is mp-invariant if and only if there exists an R-GNN N such that $\forall G \in \mathcal{CG}_B \quad N(G) = F(G)$. Furthermore, if F is computable in time $T(n)$ and space $S(n)$, then N uses time $O(T(n)) + \text{poly}(n)$ and space $O(S(n)) + \text{poly}(n)$.*

So far, R-GNNs can only compute mp-invariant functions. For connected graphs, we can break the invariance by introducing *random node initialization* (RNI) (Abboud et al. 2021), that is, augmenting the initial feature of each node by a random number $r \sim U(0, 1)$. GNNs with RNI describe a randomized algorithm, yet this randomized algorithm, or the random variable it computes, still satisfies the usual equivariance condition i.e. its output probability-distribution is identical for every two equivariant vertices (see (Abboud et al. 2021)). We say that an R-GNN N with RNI computes a function $F : \mathcal{G}_B \rightarrow \mathcal{Z}_B$ on a graph $G \in \mathcal{G}_B$ and node $v \in V(G)$ if and only if: $\Pr(N(G, v) = F(G, v)) \geq p$ for some $\frac{1}{2} < p$, and $N(G, v) \neq F(G, v) \Leftrightarrow N(G, v) = \text{null-value}$. By repeatedly running N we can boost $\Pr(N(G, v) = F(G, v))$ arbitrarily close to 1.

Corollary 5.2. *Let $\mathcal{CG}_B \subset \mathcal{G}_B$ be the subset of connected graphs in \mathcal{G}_B , and let $F : \mathcal{CG}_B \rightarrow \mathcal{Z}_B$ be computable in time $T(n)$ and space $S(n)$. Then, there exists an R-GNN N with RNI, such that F is computable by N . Furthermore, N uses time $O(T(n)) + \text{poly}(n)$, space $O(S(n)) + \text{poly}(n)$, and $O(n \log n)$ random bits.*

Another variant of our main result concerns a common extension of GNNs which is the addition of *global sum-aggregation* (a.k.a. *global sum*; *virtual nodes*), i.e. a sum-aggregation of the features of all nodes, as a third input to the combine MLP (Gilmer et al. 2017; Barceló et al. 2020). Adapting Definition 4.1, let $N = (A, F)$ be a GNN with global sum, then F has I/O dimensions $3d; d$, and

$$N^{(i+1)}(G, v) := F \left(N^{(i)}(G, v), \sum_{w \in N_G(v)} N^{(i)}(G, w), \sum_{w \in V(G)} N^{(i)}(G, w) \right)$$

Note that here we do not need the size of the graph as an input, since it can easily be computed using global sum. Instead of mp-invariance, R-GNNs with global sum satisfy a different invariance that we call *WL-invariance*. It is based on a variant of the Color Refinement algorithm, the 1-dimensional Weisfeiler-Leman algorithm, that captures the additional global information obtained through global sum. See the appendix and (Grohe 2021) for more details.

Theorem 5.3. *Let $F : \mathcal{G}_B \rightarrow \mathcal{Z}_B$ be a computable feature transformation. Then F is WL-invariant if and only if there is an R-GNN with global aggregation that computes F . Furthermore, if F is computable in time $T(n)$ and space $S(n)$, then the R-GNN uses time $O(T(n)) + \text{poly}(n)$ and space $O(S(n)) + \text{poly}(n)$.*

This theorem also has a version for graph embeddings. Here it is not restricted to connected graphs, since global aggregation provides access to all connected components.

Corollary 5.4. *Let $F : \mathcal{G}_B \rightarrow \mathcal{B}$ be computable. Then F is WL-invariant if and only if it is computable by an R-GNN. Furthermore, if F is computable in time $T(n)$ and space $S(n)$, then the R-GNN uses time $O(T(n)) + \text{poly}(n)$ and space $O(S(n)) + \text{poly}(n)$.*

6 Concluding Remarks

We prove that recurrent graph neural networks can emulate any message-passing algorithm, with only a polynomial time and space overhead. Thus recurrent graph neural networks are universal for message-passing algorithms, or computable mp-invariant functions. Note that our theorem is not an approximation theorem; by focusing on computable functions, we can actually construct GNNs computing the functions exactly. By adding randomization, we can even overcome the limitation to mp-invariant functions.

Our time-complexity analysis for the reduction states "polynomial overhead", and it is not difficult to extract an upper bound of $O(n^{10}k^2)$ from our proofs. It will be useful to know whether our reduction can be improved so to have a lower complexity, and to have a lower bound for any reduction from computable mp-invariant functions to R-GNNs. Ideally, a tight bound would be constructively proven. These remain open.

Acknowledgments

During the work on this paper, Eran Rosenbluth was funded by the European Union (ERC, SymSim, 101054974) and German Research Council (DFG), RTG 2236 (UnRAVeL). Martin Grohe was funded by the European Union (ERC, SymSim, 101054974).

References

- Aamand, A.; Chen, J.; Indyk, P.; Narayanan, S.; Rubinfeld, R.; Schiefer, N.; Silwal, S.; and Wagner, T. 2022. Exponentially improving the complexity of simulating the Weisfeiler-Lehman test with graph neural networks. *Advances in Neural Information Processing Systems*, 35: 27333–27346.
- Abboud, R.; Ceylan, İ. İ.; Grohe, M.; and Lukasiewicz, T. 2021. The Surprising Power of Graph Neural Networks with Random Node Initialization. In Zhou, Z.-H., ed., *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 2112–2118.
- Ahvonnen, V.; Heiman, D.; Kuusisto, A.; and Lutz, C. 2024. Logical characterizations of recurrent graph neural networks with reals and floats. *Advances in Neural Information Processing Systems*, 37: 104205–104249.
- Barceló, P.; Kostylev, E. V.; Monet, M.; Pérez, J.; Reutter, J. L.; and Silva, J. P. 2020. The Logical Expressiveness of Graph Neural Networks. In *8th International Conference on Learning Representations (ICLR 2020)*. OpenReview.net.
- Bravo, C.; Kozachinskiy, A.; and Rojas, C. 2024. On dimensionality of feature vectors in MPNNs. In *Forty-first International Conference on Machine Learning*.
- Bresson, X.; and Laurent, T. 2018. Residual Gated Graph ConvNets.
- Cardon, A.; and Crochemore, M. 1982. Partitioning a Graph in $O(|A| \log_2 |V|)$. *Theor. Comput. Sci.*, 19: 85–98.
- Chen, Z.; Villar, S.; Chen, L.; and Bruna, J. 2019. On the equivalence between graph isomorphism testing and function approximation with gnns. *Advances in neural information processing systems*, 32.
- Gallicchio, C.; and Micheli, A. 2010. Graph echo state networks. In *Proceedings of the IEEE International Joint Conference on Neural Networks*.
- Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR.
- Grohe, M. 2021. The Logic of Graph Neural Networks. In *36th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2021, Rome, Italy, June 29 - July 2, 2021*, 1–17. IEEE.
- Grohe, M. 2023. The Descriptive Complexity of Graph Neural Networks. In *Proceedings of the 38th Annual ACM/IEEE Symposium on Logic in Computer Science*.
- Grohe, M.; and Rosenbluth, E. 2024. Are Targeted Messages More Effective? In *Proceedings of the 39th Annual ACM/IEEE Symposium on Logic in Computer Science*, 1–14.
- Grohe, M.; Standke, C.; Steegmans, J.; and den Bussche, J. V. 2025. Query Languages for Neural Networks. In Roy, S.; and Kara, A., eds., *28th International Conference on Database Theory, ICDT 2025, March 25-28, 2025, Barcelona, Spain*, volume 328 of *LIPICS*, 9:1–9:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.
- Hornik, K.; Stinchcombe, M.; and White, H. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5): 359–366.
- Kiefer, S.; Schweitzer, P.; and Selman, E. 2022. Graphs Identified by Logics with Counting. *ACM Trans. Comput. Log.*, 23(1): 1:1–1:31.
- Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Li, Y.; Zemel, R.; Brockschmidt, M.; and Tarlow, D. 2016. Gated Graph Sequence Neural Networks. In *Proceedings of ICLR’16*.
- Morgan, H. 1965. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2): 107–113.
- Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W. L.; Lenssen, J. E.; Rattan, G.; and Grohe, M. 2019. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 4602–4609.
- Otto, M. 1997. Canonization for Two Variables and Puzzles on the Square. *Ann. Pure Appl. Log.*, 85(3): 243–282.
- Pfluger, M.; Cucala, D. T.; and Kostylev, E. V. 2024. Recurrent Graph Neural Networks and Their Connections to Bisimulation and Logic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14608–14616.
- Rosenbluth, E.; Tönshoff, J.; and Grohe, M. 2023. Some might say all you need is sum. *arXiv preprint arXiv:2302.11603*.
- Rosenbluth, E.; Tönshoff, J.; Ritzert, M.; Kisim, B.; and Grohe, M. 2024. Distinguished In Uniform: Self-Attention Vs. Virtual Nodes. In *The Twelfth International Conference on Learning Representations*.
- Scarselli, F.; Gori, M.; Tsai, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80.
- Selsam, D.; Lamm, M.; Benedikt, B.; Liang, P.; de Moura, L.; Dill, D. L.; et al. 2016. Learning a SAT Solver from Single-Bit Supervision. In *International Conference on Learning Representations*.
- Siegelmann, H. T.; and Sontag, E. D. 1992. On the computational power of neural nets. In *Proceedings of the fifth annual workshop on Computational learning theory*, 440–449.
- Tönshoff, J.; Kisim, B.; Lindner, J.; and Grohe, M. 2023. One model, any CSP: graph neural networks as fast global search heuristics for constraint satisfaction. In *Proceedings of the*

Thirty-Second International Joint Conference on Artificial Intelligence, 4280–4288.

Weisfeiler, B.; and Leman, A. 1968. The reduction of a graph to canonical form and the algebra which appears therein. *NTI, Series 2*. English translation by G. Ryabov available at https://www.iti.zcu.cz/wl2018/pdf/wl_paper_translation.pdf.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How Powerful are Graph Neural Networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

A Color Refinement and Weisfeiler Leman

The Representation Of a Color

We introduce a succinct representation of polynomial size, see Figure 3(c). Every $c \in \text{MPC}_{n,k}^{(t)}$ will be represented by a directed acyclic graph (dag) $D(c)$ with labelled edges and labelled leaves and possibly with multiple edges between the same nodes. $D(c)$ will have $t+1$ levels (numbered $0, \dots, t$), each with at most n vertices. All edges will go from some level $i \geq 1$ to level $i-1$. Edges will be labelled by natural numbers in the range $0, \dots, n-1$. Every vertex v on a level i will represent some color $\gamma(v) \in \text{MPC}_{n,k}^{(i)}$ in such a way that the mapping γ is injective, that is, every $d \in \text{MPC}_{n,k}^{(i)}$ is represented by at most one node.

We say that a color d is an element of a color c and write $d \in c$ if $c = (d_0, \{d_1, \dots, d_k\})$ and $d \in \{d_0, \dots, d_k\}$. We construct the dag $D(c)$ for $c \in \text{MPC}_{n,k}^{(t)}$ inductively as follows.

- Level t consists of a single node u , and we let $\gamma(u) = c$.
- Suppose that for some $i \in [t]$ we have defined level i and that C_i is the set of all colors represented by a node on level i . Note that $|C_i| \leq n$, because at most n distinct colors can appear in a graph of order n .

We let C_{i-1} be the set of all elements of colors in C_i . For every $d \in C_{i-1}$ we introduce a node v_d on level $i-1$ representing d , that is, $\gamma(v_d) := d$.

To define the edges, let v be a node on level i with $\gamma(v) = (d_0, \{d_1, \dots, d_k\})$. We add an edge with label 0 from v to the unique node v' on level $i-1$ with $\gamma(v') = d_0$. Moreover, for every $d \in C_{i-1}$ such that the multiplicity of d in the multiset $\{d_1, \dots, d_k\}$ is $\ell \geq 1$ we add an edge labelled ℓ from v to the node v'' on level $i-1$ with $\gamma(v'') = d$.

- We still need to define the labels of the leaves, that is, the nodes on level 0. Every node v on level 0 represents an initial color $\gamma(v) \in \text{MPC}_{n,k}^{(0)}$. Such a color is a feature of a graph $G \in \mathcal{G}_{B_k}$, that is, a bitstring of length k , and we label v by this bitstring.

This completes the description of $D(c)$.

Example A.1. Consider the graph in Figure 2(a). The colors in the figure represent the colors reached after two rounds of Color Refinement. Let us consider one of the colors, say blue, and explain how the (blue) dag representing this color is constructed. The graph has no features, so all nodes get the same initial color, corresponding to the unique bottom node of the dag. Colors reached after the first round of Color Refinement correspond to the degrees of the nodes. In our blue dag, we need two of these colors, "degree 2" and "degree 3". The left node of level 1 of the blue dag represents "degree 2"; it has an edge labeled 2 to the unique node (color) on level 0. The right node of level 1 represents "degree 3"; it has an edge labeled 3 to the unique node on level 0. In addition, both nodes on level 1 have an edge labeled 0 to the unique node on level 0, indicating that in the previous round they both had the color represented by that

node. On level 2 we have just one node, representing the color blue. A blue node has one neighbor of degree 2 and one neighbor of degree 3. Hence the top node in the dag has an edge labeled 1 to the node on level 1 representing "degree 2" and an edge labeled 1 to the node on level 1 representing "degree 3". Moreover, a blue node has degree 2 itself. Hence it has an edge labeled 0 to the node on level 1 representing "degree 2".

Observe that $|D(c)| \leq (t+1)n$ and that for $c, d \in \text{MPC}_{n,k}^{(t)}$ we have $D(c) = D(d) \iff c = d$. Furthermore, it is easy to see that given a graph $G \in \mathcal{G}_{B_k}$ of order n , a node $v \in V(G)$, and a $t \in \mathbb{N}$, we can compute $D(\text{mpc}_G^{(t)}(v))$ in time polynomial in k, n, t . To do this, we construct a dag simultaneously representing all colors appearing in a graph. Once we have this, we can construct a dag only representing a single color by deleting unreachable nodes. We construct the dag level by level in a bottom up fashion, maintaining pointers from the nodes of the graph to the nodes of the dag representing their color at the current level. At each level, we need to sweep through all nodes of the graph and all their incident edges, which overall requires time $O(n+m)$, where m is the number of edges of the graph. Thus overall, to construct t levels of the dag we need time $O(t(n+m))$.

Proof of Lemma 3.2

Before we prove the lemma, we need to make some additional remarks on Color Refinement. Consider a graph G of order $n := |G|$. After each refinement round, we obtain a partition $\Pi_G^{(t)}$ of $V(G)$ into color classes $\{v \in V(G) \mid \text{mpc}_G^{(t)}(v) = c\}$, for $c \in \text{MPC}^{(t)}$. As the partition $\Pi_G^{(t+1)}$ refines the partition $\Pi_G^{(t)}$ and as $\Pi_G^{(t+1)} = \Pi_G^{(t)}$ implies $\Pi_G^{(t+2)} = \Pi_G^{(t+1)}$, for all t , there is a $t \leq n-1$ such that $\Pi_G^{(t+s)} = \Pi_G^{(t)}$ for all s . We call $\Pi^{(t)}$ the coarsest stable partition of G . However, note that even if $\Pi^{(t+1)} = \Pi^{(t)}$ we have $\text{mpc}_G^{(t+1)}(v) \neq \text{mpc}_G^{(t)}(v)$ for all v in $V(G)$, and potentially $\text{mpc}_G^{(t+1)}(v)$ contains relevant information not captured by $\text{mpc}_G^{(t)}(v)$. We will see later (Lemma A.5) that for $t \geq 2n$, this will no longer happen. For this reason, we define the final color $\text{mpc}_G(v)$ to be $\text{mpc}_G^{(2n)}(v)$.

It is known that the coarsest stable partition of G can be computed in time $O(n^2 \log n)$ (Cardon and Crochemore 1982). This does not mean that we can efficiently compute the colors $\text{mpc}_G(v)$, which may be exponentially large, but, as explained earlier, we can compute $D(\text{mpc}_G^{(t)}(v))$ in time polynomial in k, n, t .

The following lemma is just restating Lemma 3.2.

Lemma A.2. Let $F : \mathcal{G}_B \rightarrow \mathcal{Z}_B$ be computable. Then the following are equivalent.

1. F is mp-invariant.
2. There is an algorithm that computes $F(G, v)$ from the color $\text{mpc}_G(v)$. More precisely, there is an algorithm that, given $D(\text{mpc}_G(v))$ for some graph $G \in \mathcal{G}_B$ and $v \in V(G)$, computes $F(G, v)$.

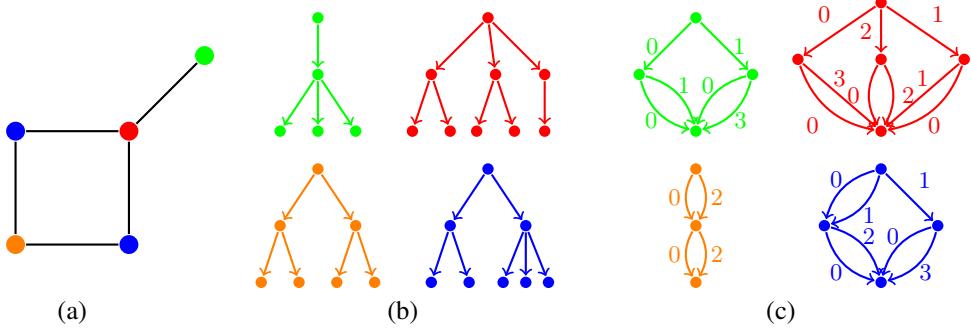


Figure 3: Two rounds of Color Refinement on a graph G shown in (a). Colors can be represented as trees (b) or dags (c). The actual colors in the figure illustrate the coloring reached after two steps (they do not indicate initial features of the nodes).

Furthermore, if F is computable in time $T(n)$ then algorithm in (2) can be constructed to run in time $T(n) + \text{poly}(n)$, and conversely, if the algorithm in (2) runs in time $T(n)$ then F is computable in time $T(n) + \text{poly}(n)$.

The proof of this lemma relies on result due to (Otto 1997). First of all, Otto observed that the complete message-passing information of a graph can be represented by a sketch of the graph.² We say that an ℓ -dimensional sketch, for some $\ell \geq 1$, is a triple $S = (A, b, c)$ where $A = (A_{ij})_{i,j \in [\ell]} \in \mathbb{N}^{\ell \times \ell}$, and $b = (b_1, \dots, b_\ell) \in \mathbb{N}_{>0}^\ell$ and $c = (c_1, \dots, c_\ell) \in \mathcal{B}^\ell$. With every graph $G \in \mathcal{G}_B$, we associate a sketch $S(G) = (A, b, c)$ as follows: suppose that the color classes of the coarsest stable coloring of G are V_1, \dots, V_ℓ . Then for all $i \in [\ell]$ we let $b_i := |V_i|$ and $c_i := Z(G, v)$ for all $v \in V_i$ (using the fact that $\text{mpc}_G(v) = \text{mpc}_G(w)$ implies $Z(G, v) = Z(G, w)$). Moreover, for all $i, j \in [\ell]$ we let A_{ij} be the number of neighbors that vertices in V_i have in V_j , that is, $A_{ij} := |N_G(v) \cap V_j|$ for $v \in V_i$ (using the fact that $\text{mpc}_G(v) = \text{mpc}_G(w)$ implies $|N_G(v) \cap V_j| = |N_G(w) \cap V_j|$ for all j). This definition depends on the order V_1, \dots, V_ℓ in which we enumerate the color classes. But it is easy to define a canonical order of these classes (see (Otto 1997; Kiefer, Schweitzer, and Selman 2022)), and this is the order we always use. The following lemma shows that the sketches capture the complete Color Refinement information. We say that Color Refinement distinguishes two graphs, if

$$\{\{\text{mpc}_G^{(t)}(v) \mid v \in V(G)\}\} \neq \{\{\text{mpc}_{G'}^{(t)}(v) \mid v \in V(G')\}\}$$

for some $t \in \mathbb{N}$.

Lemma A.3 ((Otto 1997)). For all $G, H \in \mathcal{G}_B$ the following are equivalent:

1. $S(G) = S(H)$;
2. Color Refinement does not distinguish G and H ;
3. $\{\{\text{mpc}_G^{(t)}(v) \mid v \in V(G)\}\} = \{\{\text{mpc}_H^{(t)}(w) \mid w \in V(H)\}\}$ for some $t \geq \max\{|G|, |H|\}$.

²In (Otto 1997; Kiefer, Schweitzer, and Selman 2022), our sketches are called C^2 -invariants, referring to a logic C^2 that is closely related to the Color Refinement algorithm.

The main result of (Otto 1997) is that we can retrieve a graph from its sketch in polynomial time, that is, there is a polynomial time algorithm that, given a sketch S , decides if there is a graph $G \in \mathcal{G}_B$ such that $S(G) = S$ and if there is computes such a graph G . Note that the graph the algorithm computes give a sketch $S(G)$ is not necessarily G , but some graph G' with $S(G') = S(G)$.

Kiefer, Selman, and Schweitzer (Kiefer, Schweitzer, and Selman 2022) gave a simpler proof of Otto's result. The core of their proof is the following lemma, which we will also use here. An ℓ -dimensional sketch $S = (A, b, c)$ is *realisable* if there is a graph G and a partition V_1, \dots, V_ℓ of $V(G)$ such that for all $i \in [\ell]$ we have $b_i = |V_i|$ and $c_i = Z(G, v)$ for all $v \in V_i$, and for all $i, j \in [\ell]$, each vertex $v \in V_i$ has A_{ij} neighbours in V_j . We say that G realises S via V_1, \dots, V_ℓ .

Note that this does not mean that $S(G) = S$, because the partition V_1, \dots, V_ℓ may be a refinement of the coarsest stable partition of G . For example, for every graph $G \in \mathcal{G}_B$ with vertex set $V(G) = \{v_1, \dots, v_n\}$ we can define a trivial n -dimensional sketch $S = (A, b, c)$ by letting $A_{ij} := 1$ if $v_i v_j \in E(G)$ and $A_{ij} = 0$ otherwise (that is, A is the adjacency matrix of G), $b_i := 1$ for all i , and $c_i := Z(G, v_i)$. It can be shown that $S(G)$ is the unique sketch of minimum dimension that is realised by G .

Lemma A.4 ((Kiefer, Schweitzer, and Selman 2022)). A sketch $S = (A, b, c)$ is realisable if and only if the following conditions are satisfied:

- (i) $b_i \cdot A_{ii}$ is even for all $i \in [\ell]$;
- (ii) $b_i A_{ij} = b_j A_{ji}$ for all $i, j \in [\ell]$.

Furthermore, there is a polynomial time algorithm that, given a sketch S , decides if S realisable and computes a graph G and a partition V_1, \dots, V_ℓ of $V(G)$ realising S if S is realisable.

We need versions of these lemmas on the node level. An ℓ -dimensional weak sketch is a pair (A, c) where $A = (A_{ij})_{i,j \in [\ell]} \in \mathbb{N}^{\ell \times \ell}$ and $c = (c_1, \dots, c_\ell) \in \mathcal{B}^\ell$. An ℓ -dimensional weak node sketch is a triplet (A, c, k) , where (A, c) is an ℓ -dimensional weak sketch and $k \in [\ell]$. The weak node sketch of a connected rooted graph (G, v) , where $G \in \mathcal{G}_B$ and $v \in V(G)$, is the weak node

sketch $S_w(G, v)$ defined as follows: suppose that the MPC-equivalence classes of G are V_1, \dots, V_ℓ (in canonical order), $v \in V_k$, and $S(G) = (A, \mathbf{b}, \mathbf{c})$. Then $S_w(G, v) = (A, \mathbf{c}, k)$. We call this the “weak” sketch because we drop the information about the sizes of the equivalence classes, represented by the vector \mathbf{b} from the sketch. The *weak node sketch* of a disconnected rooted graph (G, v) is the weak node sketch $S_w(G, v) := S_w(G_v, v)$, where G_v is the connected component of v in G . The idea behind this definition being that the weak node sketch should only contain information accessible to v by message-passing. In the following, we always denote the connected component of a vertex v in a graph G by G_v .

A weak node sketch $S = (A, \mathbf{c}, k)$ is *realisable* if there is a rooted graph (G, v) and a partition V_1, \dots, V_ℓ of $V(G_v)$ such that for all $i \in [\ell]$, $v' \in V_i$ we have $c_i = Z(G, v')$, and for all $i, j \in [\ell]$, each vertex $v' \in V_i$ has A_{ij} neighbours in V_j . We say that (G, v) *realises* S via V_1, \dots, V_ℓ .

Lemma A.5. *Let $G, H \in \mathcal{G}_B$ be connected graphs, $v \in V(G)$, $w \in V(H)$, and $t \geq 2 \max\{|G|, |H|\}$. Then the following are equivalent:*

1. *there is a weak node sketch S such that both (G, v) and (H, w) realise S ;*
2. $S_w(G, v) = S_w(H, w)$;
3. $\text{mpc}_G^{(t)}(v) = \text{mpc}_H^{(t)}(w)$;
4. $\text{mpc}_G^{(s)}(v) = \text{mpc}_H^{(s)}(w)$ for all $s \geq 1$.

Proof. The implications (2) \implies (1) and (4) \implies (3) are trivial. To prove that (3) \implies (2), assume that $\text{mpc}_G^{(t)}(v) = \text{mpc}_H^{(t)}(w)$ and consider the dag $D := D(\text{mpc}_G^{(t)}(v)) = D(\text{mpc}_H^{(t)}(w))$. Recall that D has t levels, with edges only between successive levels. For $i \in [t]$, let $U^{(i)}$ be the set of all nodes on level (i) .

Each node $u \in U^{(i)}$ represents the message-passing information $\text{mpc}_G^{(i)}(v')$ of at least one node $v' \in V(G)$ and the message-passing information $\text{mpc}_H^{(i)}(w')$ of at least one node $w' \in V(H)$. Let $V(u) \subseteq V(G)$ and $W(u) \subseteq V(H)$ be the sets of all nodes $v' \in V(G)$, $w' \in V(H)$, respectively, whose message-passing information is represented by u . Moreover, let $\text{mpc}(u) := \text{mpc}_G^{(i)}(v') = \text{mpc}_G^{(i)}(w')$ for all $v' \in V(u)$, $w' \in W(u)$.

If a node $v' \in V(G)$ is reachable from v by a path of length at most j , then $v' \in V(u)$ for some $u \in U^{(i)}$ for every $i \leq t - j$. Let $n := \max\{|G|, |H|\} \leq \lfloor t/2 \rfloor$. Since G is connected, every node is reachable by a path of length at most $n - 1$. This implies that for all $i \leq n + 1$, $(V(u))_{u \in U^{(i)}}$ is a partition of $V(G)$. Similarly, $(W(u))_{u \in U^{(i)}}$ is a partition of $V(H)$. Thus

$$\begin{aligned} \{\text{mpc}(u) \mid u \in U^{(i)}\} &= \{\text{mpc}_G^{(i)}(v') \mid v' \in V(G)\} = \\ &= \{\text{mpc}_H^{(i)}(w') \mid w' \in V(H)\}. \end{aligned}$$

Note that these are set equalities, not multiset equalities. It may well be that for $u \in U^{(i)}$ we have $|V(u)| \neq$

$|W(u)|$, and thus the multiplicities of $\text{mpc}(u)$ in the multisets $\{\{\text{mpc}_G^{(i)}(v') \mid v' \in V(G)\}\}$ and $\{\{\text{mpc}_H^{(i)}(w') \mid w' \in V(H)\}\}$ are different.

As the refinement process on G and H stabilizes in at most n iterations, $(V(u))_{u \in U^{(n)}}$ is the coarsest stable partition of G . Moreover, as $(V(u))_{u \in U^{(n+1)}}$ is still a partition of $V(G)$, we have $(V(u))_{u \in U^{(n+1)}} = (V(u'))_{u' \in U^{(n)}}$. The corresponding matching between $U^{(n+1)}$ and $U^{(n)}$ is given by the edges in D from level $n + 1$ to level n with label 0. For all $u \in U^{(n+1)}$, $u' \in U^{(n)}$, the message passing information $\text{mpc}(u)$ tells us the number of neighbours each $v \in V(u)$ has in $V(u')$. Thus for all $u, u' \in U^{(n)}$ there is a number $A_{uu'}$ such that each $v \in V(u)$ has $A_{uu'}$ neighbours in $V(u')$. Similarly, $(W(u))_{u \in U^{(n)}}$ is the coarsest stable partition of H , and for all u, u' each vertex in $W(u)$ has $A_{uu'}$ neighbours in $W(u')$; since $A_{uu'}$ only depends on $\text{mpc}(u'')$ for the node $u'' \in U^{(n+1)}$ connected to u by an edge with label 0, it is the same in both G and H .

Furthermore, for all $u \in U^{(n)}$ there is a $c_u \in \mathcal{B}$ determined by $\text{mpc}(u)$ such that $Z(G, v') = c_u$ for all $v' \in V(u)$ and $Z(H, w') = c_u$ for all $w' \in W(u)$.

Letting $A := (A_{uu'})_{u, u' \in U^{(n)}}$, $\mathbf{b} = (b_u)_{u \in U^{(n)}}$ with $b_u := |V(u)|$, $\mathbf{b}' = (b'_u)_{u \in U^{(n)}}$ with $b'_u := |W(u)|$, and $\mathbf{c} = (c_u)_{u \in U^{(n)}}$, we have $S(G) = (A, \mathbf{b}, \mathbf{c})$ and $S(H) = (A, \mathbf{b}', \mathbf{c})$.

Moreover, tracing the unique path of edges labelled 0 from the unique node $u \in U^{(t)}$ with $\text{mpc}(u) = \text{mpc}_G^{(t)}(v) = \text{mpc}_H^{(t)}(w)$ to level n , we find a node $u' \in U^{(n)}$ such that $\text{mpc}(u') = \text{mpc}_G^{(n)}(v) = \text{mpc}_H^{(n)}(w)$. Then we have

$$S_w(G, v) = (A, \mathbf{c}, u') = S_w(H, w).$$

It remains to prove the implication (1) \implies (4). Assume that $S = (A, \mathbf{c}, k)$, where $A = (A_{ij})_{i, j \in [\ell]}$ and $\mathbf{c} = (c_1, \dots, c_\ell)$ for some $\ell \in \mathbb{N}$, is a weak node sketch realised by (G, v) via the partition V_1, \dots, V_ℓ and realised by (H, w) via the partition W_1, \dots, W_ℓ .

Let $s \geq 1$. It is our goal to prove that

$$\text{mpc}_G^{(s)}(v) = \text{mpc}_H^{(s)}(w). \quad (\text{A.1})$$

Let $n := \max\{|G|, |H|\}$, and let $D_G := D(\text{mpc}_G^{(s+n)}(v))$ and $D_H := D(\text{mpc}_H^{(s+n)}(w))$. For every $i \in [s+n]$, let $D_G^{(i)}$ be the restriction of D_G to the first i levels, and let $D_H^{(i)}$ be the restriction of D_H to the first i levels. Let $U_G^{(i)}, U_H^{(i)}$ be the sets of nodes on level i in the respective graph. Every $u \in U_G^{(i)}$ represents some color $\text{mpc}_G(u)$ such that there is a $v' \in V(G)$ with $\text{mpc}_G^{(i)}(v') = \text{mpc}_G(u)$. We let $V_u^{(i)}$ be the set of all such v' . Then the sets $V_u^{(i)}$ for $u \in U_G^{(i)}$ are mutually disjoint. Arguing as above, we see that for $i \leq s$ the sets $V_u^{(i)}$ for $u \in U_G^{(i)}$ form a partition of $V(G)$. Similarly, for every $u \in U_H^{(i)}$ there is some color $\text{mpc}_H(u)$ and a vertex $w' \in V(H)$ with $\text{mpc}_H^{(i)}(w') = \text{mpc}_H(u)$. We let $W_u^{(i)}$ be the set of all such w' . Then for $i \leq s$, the sets $W_u^{(i)}$ for $u \in U_H^{(i)}$ form a partition of $V(H)$.

By induction, we shall prove that for all $i \in [s]$, the following conditions are satisfied.

$$(i) D_G^{(i)} = D_H^{(i)}.$$

In particular, this implies $U_G^{(i)} = U_H^{(i)} =: U^{(i)}$ and $\text{mpc}_G(u) = \text{mpc}_H(u) =: \text{mpc}(u)$ for all $u \in U^{(i)}$.

- (ii) The partition $(V_i)_{i \in [\ell]}$ refines the partition $(V_u^{(i)})_{u \in U^{(i)}}$.
- (iii) The partition $(W_i)_{i \in [\ell]}$ refines the partition $(W_u^{(i)})_{u \in U^{(i)}}$.
- (iv) For every $u \in U^{(i)}$ and $j \in [\ell]$ we have $V_j \subseteq V_u^{(i)} \iff W_j \subseteq W_u^{(i)}$.

Observe that (i) implies (A.1).

For the base step, we recall that $\text{mpc}_G^{(1)}(v') = Z(G, v') \in \mathcal{B}$, and that for every $c \in \mathcal{B}_1$, we have

$$\{v' \in V(G) \mid Z(G, v') = c\} = \bigcup_{i \in [\ell] \text{ with } c_i=c} V_i,$$

$$\{w' \in V(H) \mid Z(H, w') = c\} = \bigcup_{i \in [\ell] \text{ with } c_i=c} W_i.$$

Thus both $D_G^{(1)}$ and $D_H^{(1)}$ consist of isolated nodes u_1, \dots, u_ℓ with $\text{mpc}_G(u_i) = \text{mpc}_H(u_i) = c_i$, and we have $V_{u_i}^{(1)} = \bigcup_{i \in [\ell] \text{ with } c_i=c} V_i$ and $W_{u_i}^{(1)} = \bigcup_{i \in [\ell] \text{ with } c_i=c} W_i$. This implies assertions (i)–(iv) for $i = 1$.

For the inductive step $i \rightarrow i + 1$, assume that we have proved (i)–(iv) for i . To prove it for $i + 1$, we shall prove that for all $j \in [\ell]$ there is a color $c_j^{(i+1)}$ such that for all $v' \in V_j$ it holds that $\text{mpc}_G^{(i+1)}(v') = c_j^{(i+1)}$ and for all $w' \in W_j$ it holds that $\text{mpc}_H^{(i+1)}(w') = c_j^{(i+1)}$. Assertions (i)–(iv) for $i + 1$ follow.

Let $j \in [\ell]$. We need to prove that for all $u \in U^{(i)}$ there is a $a_{ju} \geq 0$ such that all $v' \in V_j$ have a_{ju} neighbours in $V_u^{(i)}$ and all $w' \in W_j$ have a_{ju} neighbours in $W_u^{(i)}$. So let $u \in U^{(i)}$. By the induction hypothesis, there are $j_1, \dots, j_k \in [\ell]$ such that $V_u^{(i)} = V_{j_1} \cup \dots \cup V_{j_k}$ and $W_u^{(i)} = W_{j_1} \cup \dots \cup W_{j_k}$. We let $a_{ju} := A_{j_1j_1} + \dots + A_{j_kj_k}$. As every $v' \in V_j$ has $A_{j_jj_p}$ neighbours in V_{j_p} , it has a_{ju} neighbours in $V_u^{(i)}$. Similarly, every $w' \in W_j$ has a_{ju} neighbours in $W_u^{(i)}$. \square

Corollary A.6. Let $G, H \in \mathcal{G}_{\mathcal{B}}$ be graphs and $v \in V(G), w \in V(H)$. Then the following are equivalent:

1. $|G| = |H|$ and there is a weak node sketch S such that both (G, v) and (H, w) realise S ;
2. $\text{mpc}_G(v) = \text{mpc}_H(w)$;

Lemma A.7. There is a polynomial-time algorithm that, given the dag $D(\text{mpc}_G(v))$ for some connected graph G and node $v \in V(G)$, computes the weak node sketch $S_w(G, v)$.

Proof. The proof of the implication (3) \implies (2) of Lemma A.5 describes how to construct $S_w(G, v)$ from $D(\text{mpc}_G^{(t)}(v))$ (without ever seeing (G, v)). It is easy to see that this construction can be turned into a polynomial time algorithm. \square

The next lemma is a node-level variant of Lemma A.4.

Lemma A.8. There is a polynomial-time algorithm that, given a weak node sketch S and an $n \geq 1$ in unary, decides if there is a rooted graph (G, v) of order n realizing S and computes such a graph if there is.

Proof. Let $S = (A, \mathbf{c}, k)$, where $A = (A_{ij})_{i,j \in [\ell]} \in \mathbb{N}^{\ell \times \ell}$, $\mathbf{c} = (c_1, \dots, c_\ell) \in \mathcal{B}^\ell$, and $k \in [\ell]$. Observe that S is not realisable if the graph $H := ([\ell], \{ij \mid A_{ij} > 0\})$ is disconnected, because in any realisation (G, v) of S the matrix carries information about the connected component G_v . So we first check if the graph H is connected and reject if it is not.

Observe next that $S = (A, \mathbf{c}, k)$ is realisable by a connected graph of order at most n if and only if there is a tuple $\mathbf{b} \in \mathbb{N}_{>0}^\ell$ such that $\sum_{i=1}^\ell b_i \leq n$ and the sketch (A, \mathbf{b}, c) is realisable. To find such a tuple \mathbf{b} , we need to satisfy $\sum_{i=1}^\ell b_i \leq n$ and conditions (i) and (ii) of Lemma A.4, which amounts to solving a system of linear equations in the variables b_i . If there is no solution, then S is not realisable by connected graph of order at most n . If we have found a solution \mathbf{b} , using Lemma A.4 we compute a connected graph G' and a partition V_1, \dots, V_ℓ realising $(A, \mathbf{b}, \mathbf{c})$. We pick an arbitrary node $v \in V_k$. Then (G', v) realises S . However, we may have $|G'| < n$. If that is the case, we extend G' by $n - |G'|$ isolated vertices to obtain a graph G . Then (G, v) still realises S , and we have $|G| = n$. \square

Corollary A.9. There is a polynomial-time algorithm that, given the dag $D(\text{mpc}_G(v))$ for some graph G and node $v \in V(G)$, computes a graph G' of order $|G'| = |G|$ and node $v' \in V(G')$ such that $\text{mpc}_{G'}(v') = \text{mpc}_G(v)$.

Proof. Let $D = D(\text{mpc}_G(v))$ for some graph G and node $v \in V(G)$, and let $n := |G|$. Since $\text{mpc}_G(v) = \text{mpc}_G^{(2n)}(v)$, the dag D has $2n + 1$ levels, and hence we can compute n from D . Using Lemma A.7 we compute $S := S_w(G, v)$.

Using Lemma A.8, we compute a rooted graph (G', v') of order $|G'| = n$ realising $S_w(G, v)$. It follows from Corollary A.6 that $\text{mpc}_{G'}(v') = \text{mpc}_G(v)$. \square

Proof of Lemma A.2. The implications (2) \implies (1) is trivial. We need to prove the converse

Suppose that F is mp-invariant. Then given $D(\text{mpc}_G^{(t)}(v))$, using Corollary A.9, we compute a graph $G' \in \mathcal{G}_{\mathcal{B}}$ and a node $v' \in V(G')$ such that $\text{mpc}_{G'}^{(t)}(v') = \text{mpc}_G^{(t)}(v)$ and $|G'| = |G|$. Then we compute $F(G', v')$. Since F is CR-invariant, this is equal to the desired $F(G, v)$.

Since the algorithm of Corollary A.9 runs in polynomial time, the additional statement on the running time follows. \square

A Version for Graph-Level Function

In general, the message-passing color of a node does not determine the complete Color Refinement information of a graph. The following example shows that there are graphs G, H of the same size and nodes $v \in V(G), w \in V(H)$

such that $\text{mpc}_G(v) = \text{mpc}_H(w)$ and yet Color Refinement distinguishes G and H .

Example A.10. Let G be a cycle of length 4, and let H be the disjoint union of a triangle and an isolated vertex. Then clearly Color Refinement distinguishes G and H . Now let $v \in V(G)$ be arbitrary, and let w be a vertex of the triangle of H . Then $\text{mpc}_G^{(t)}(v) = \text{mpc}_H^{(t)}(w)$ for all $t \in \mathbb{N}$.

However, this cannot happen for connected graphs.

Lemma A.11. Let G, H be connected graphs of order $n := |G| = |H|$, and let $v_0 \in V(G), w_0 \in V(H)$ such that $\text{mpc}_G^{(t)}(v_0) = \text{mpc}_H^{(t)}(w_0)$ for some $t \geq 2n$. Then Color Refinement does not distinguish G and H .

Proof. It follows from Lemma A.5 that (G, v_0) and (H, w_0) have the same weak node sketch, say,

$$(A, c, k) := S_w(G, v_0) = S_w(H, w_0),$$

where $A \in \mathbb{N}^{\ell \times \ell}$ and $c \in \mathcal{B}^\ell$.

Suppose that the MPC-equivalence classes of G are V_1, \dots, V_ℓ and that the MPC-equivalence classes of H are W_1, \dots, W_ℓ (both in canonical order). For every $i \in [\ell]$, let $x_i := |V_i|$ and $y_i := |W_i|$. We need to prove that for all $i \in [\ell]$ we have $x_i = y_i$.

Recall that for all distinct $i, j \in [\ell]$, every $v \in V_i$ has exactly A_{ij} neighbors in V_j . Hence the number of edges in G between V_i and V_j is $x_i A_{ji}$. By the same reasoning applied from j to i , the number of edges between V_i and V_j is also $x_j A_{ij}$. Thus

$$x_i A_{ij} = x_j A_{ji}. \quad (\text{A.2})$$

Similarly,

$$y_i A_{ij} = y_j A_{ji}. \quad (\text{A.3})$$

Claim 1. For every $i \in [\ell]$ there is an $a_i \neq 0$ such that $x_i = a_i x_k$ and $y_i = a_i y_k$.

Proof. Let $i \in [\ell]$. If $i = k$, we let $a_i := 1$, and the assertion is trivial. Otherwise, since G is connected, there is an $m \geq 1$ and a sequence i_1, \dots, i_m such that $i_1 = i$ and $i_m = k$ and $A_{i_j i_{j+1}} \neq 0$ for $1 \leq j < m$. By (A.2), we have $x_j := \frac{A_{i_j+1 i_j}}{A_{i_j i_{j+1}}} x_{j+1}$. Thus with

$$a_i := \prod_{j=1}^{m-1} \frac{A_{i_{j+1} i_j}}{A_{i_j i_{j+1}}}$$

we have $x_i = a_i x_k$. Similarly, using (A.3), we obtain $y_i = a_i y_k$. This proves the claim.

Since $\sum_{i=1}^k x_i = |G| = |H| = \sum_{i=1}^k y_i$, by the claim we thus have

$$x_k \sum_{i=1}^k a_i = y_k \sum_{i=1}^k a_i,$$

which implies $x_k = y_k$ and, again by the claim, $x_i = a_i x_k = a_i y_k = y_i$ for all i . \square

Corollary A.12. There is a polynomial-time algorithm that, given the dag $D(\text{mpc}_G(v))$ for some connected graph G and node $v \in V(G)$, computes a graph G' such that Color Refinement does not distinguish G and G' .

The Weisfeiler-Leman Algorithm

Following (Grohe 2021), we distinguish between the Color Refinement algorithm and the Weisfeiler-Leman algorithm. In the literature, this distinction is usually not made, and what we call Color Refinement here is called Weisfeiler-Leman. Regardless of the terminology, there are two different algorithms, and the difference between them is often overlooked or ignored, because in many situation it is not relevant. However, it does make a difference here. (We refer the reader to (Grohe 2021) for a discussion and an example illustrating the difference.)

Where Color Refinement collects the local message-passing information along the edges of a graph, Weisfeiler-Leman no longer restrict the information flow to edges, but considers the information flow along non-edges as well. For every $t \geq 0$ and $v \in V(G)$, we define a color $\text{wl}_G^{(t)}(v)$ inductively as follows:

- $\text{wl}_G^{(0)}(v) := \text{mpc}_G^{(0)} = Z(G, v)$.
- $\text{wl}_G^{(t+1)}(v) := (\text{wl}_G^{(t)}(v), \{\{\text{wl}_G^{(t)}(w) \mid w \in N_G(v)\}, \{\{\text{wl}_G^{(t)}(w) \mid w \in V(G) \setminus N_G(v)\}\})$.

For WL, we define the final color to be $\text{wl}_G(v) := \text{wl}_G^{(|G|)}(v)$. We define set $\text{WL}_{n,k}^{(t)}$ analogous to the corresponding set $\text{MPC}_{n,k}^{(t)}$. We can also represent the WL colors by a dag, the simplest way of doing this is to also introduce edges with negative labels to represent the multiset of colors of non-neighbours.

Although it seems that the non-local message passing in WL adds considerable power, this is actually not the case. In particular, CR and WL have exactly the same power when it comes to distinguishing graphs: it can be shown that for all G, H , CR distinguishes G, H if and only if WL distinguishes G, H (see (Grohe 2021) for a proof).

Note, however, that this is no longer the case on the node level. For example, if G is cycle of length 3 and H a cycle of length 4 then for all $v \in V(G)$ and $w \in V(H)$ it holds that $\text{mpc}_G(v) = \text{mpc}_H(w)$ and $\text{wl}_G(v) \neq \text{wl}_H(w)$. The distinguishing power of WL on the node level exactly corresponds to that of GNNs with global readout.

There is also a version of Lemma 3.2 for the WL-algorithm. A feature transformation $F : \mathcal{G}_\mathcal{B} \rightarrow \mathcal{Z}_\mathcal{B}$ is *WL-invariant* if for all graphs $G, H \in \mathcal{G}_\mathcal{B}$ and nodes $v \in V(G), w \in V(H)$, if $\text{wl}_G^{(t)}(v) = \text{wl}_H^{(t)}(w)$ for all $t \geq 1$ then $F(G, v) = F(H, w)$. Note that each mp-invariant feature transformation is also WL-invariant, but that the converse does not hold.

Lemma A.13. Let $F : \mathcal{G}_\mathcal{B} \rightarrow \mathcal{Z}_\mathcal{B}$ be computable. Then the following are equivalent.

1. F is WL-invariant.
2. There is an algorithm that computes $F(G, v)$ from $\text{wl}_G(v)$.
3. There is an algorithm that computes $F(G, v)$ from $\text{wl}_G^{(t)}(v)$ for an arbitrary $t \geq |G|$.

Furthermore, if F is computable in time $T(n)$ then algorithms in (2) and (3) can be constructed to run in time

$T(n) + \text{poly}(n)$, and conversely, if the algorithm in either (2) or (3) runs in time $T(n)$ then F is computable in time $T(n) + \text{poly}(n)$.

The proof of this lemma is similar to the proof of Lemma A.2. Instead of the weak node sketch, we work with the *node sketch* $S(G, v) = (A, b, c, k)$, where $(A, b, c) = S(G)$ is the sketch of G and k the index of the class of the coarsest stable partition that contains v .

B Main Result

For a binary string $\mathcal{B}_k \ni x = b_1, \dots, b_k$ define its integer value $\text{B2I}(x) := \sum_{i \in [k]} b_i 2^{i-1}$, and for binary strings $x_1, x_2 \in \mathcal{B}$ define their binary sum $x_1 + x_2 := \text{B2I}^{-1}(\text{B2I}(x_1) + \text{B2I}(x_2))$. When clear from the context, we may refer to $x \in \mathbb{N}$ while meaning its binary representation $\text{B2I}^{-1}(x)$.

Definition B.1. A ReLU-activated Multilayer Perceptron (MLP) $F = (l_1, \dots, l_m), l_i = (w_i, b_i)$, of I/O dimensions d_{in}, d_{out} , and depth m , is a sequence of rational matrices w_i and bias vectors b_i such that

$$\begin{aligned} \dim(w_1)(2) &= d_{in}, \dim(w_m)(1) = d_{out}, \\ \forall i > 1 \quad \dim(w_i)(2) &= \dim(w_{i-1})(1), \\ \forall i \in [m] \quad \dim(b_i) &= \dim(w_i)(1) \end{aligned}$$

It defines a function $f_F(x) :=$

$$\text{ReLU}(w_m(\dots \text{ReLU}(w_2(\text{ReLU}(w_1(x) + b_1)) + b_2)\dots) + b_m)$$

When clear from the context, we may use $F(x)$ to denote $f_F(x)$.

Theorem B.2. There exists a feature transformation $F : \mathcal{G}_{\mathcal{B}} \rightarrow \mathcal{Z}_{\mathcal{B}}$ such that for every message-passing algorithm \mathcal{A} where the graph-order input is omitted, there exist $G \in \mathcal{G}_{\mathcal{B}}, v \in V(G)$ for which $\mathcal{A}(G, v) \neq F(G, v)$.

Proof. Consider the function F defined by $F(G, v) = 1$ if v is contained in a cycle of G and $F(G, v) = 0$ otherwise. It is not hard to see that F is mp-invariant.

Suppose that there is a message-passing algorithm \mathcal{A} , where the graph-order input is omitted, that computes F . Consider the computation of \mathcal{A} on a cycle. Regardless of the length of the cycle, the computation will be the same, because for all cycle C, C' , all nodes $v \in V(C), v' \in V(C')$, and all $t \in \mathbb{N}$ it holds that $\text{mpc}_{C'}^{(t)}(v) = \text{mpc}_{C'}^{(t)}(v')$. Hence there is an $I \in \mathbb{N}$ such that for all cycles C and nodes $v \in V(C)$ we have $I_v = I$, where $I_v \in \mathbb{N}$ is the finishing iteration of \mathcal{A} on C, v . That is, the computation terminates after I rounds and returns the value $F(C, v) = 1$.

Now consider a long path P of even length $\geq 2I$ and let w be the middle node of this path. As the neighborhood of radius I of w is identical to the neighborhood to a node v on a cycle C of length $\geq 2I$, we have $\mathcal{A}^{(i)}(P, w) = \mathcal{A}^{(i)}(C, v)$ for all $i \leq I$. Hence $I_w = I$ and $\mathcal{A}(P, w) = \mathcal{A}(C, v) = 1$, hence $\mathcal{A}(P, w) \neq F(P, w)$. \square

The next three lemmas state the reductions to intermediate models. Figure 4 illustrates the composition of these reductions from a recurrences perspective.

Lemma 4.5. Let $C = (M)$ be an MPC-GA, then there exist $A \in \mathcal{B}, f : \mathcal{B}^2 \rightarrow \mathcal{B}^2$ such that $C' = (A, f)$ is an MP-LGA and $\forall G \in \mathcal{G}_{\mathcal{B}} \forall v \in V(G) C(G, v) = C'(G, v)$. Furthermore, C' incurs polynomial time and space overhead.

Proof. The idea is to construct $\text{mpc}_G(v)$ step by step in the first $2|G|$ applications of f . Then, in the $(2|G| + 1)$ application, compute the function determined by M , on the constructed $\text{mpc}_G(v)$. The state of the computation remembers the last iteration-number t (up to $2|G| + 1$), the graph size, and $\text{mpc}_G^{(t)}(v)$. The required sum-of-messages length limit is implied by the dag-construction description in Section A. For an encoding of a triplet of binary strings $x = \theta(b_1, b_2, b_3), b_i \in \mathcal{B}$, define:

$x.t := \text{B2I}(b_1)$, representing the iteration-number part

$x.s := \text{B2I}(b_2)$, representing the graph size part

$x.d := b_3$, representing the $\delta(\text{mpc}_G^{(x.t)}(v))$ part.

Let $\delta(c_1), \dots, \delta(c_l), \forall i \in [l] c_i \in \text{MPC}_{|G|, k}^{(t)}$ be the dag encodings of a vertex v and its neighbors colors after t Color Refinement iterations, for some k, t . We define the operation of combining these encodings into the dag encoding of the next-iteration color of v .

$$\text{dc}(\delta(c_1), \{\delta(c_2), \dots, \delta(c_l)\}) := \delta(c_1, \{(c_2), \dots, (c_l)\})$$

We define (A, f) as follows: $A = 0$, representing an initial iteration-number of zero. $f(x_1, x_2) :=$

$$\begin{cases} (\theta(x_1.t + 1, x_1.s, \delta(x_1.d, x_2)), \text{dc}(x_1.d, x_2)) & x_1.t \leq x_1.s \\ (\theta(x_1.t + 1, x_1.s, x_1.d), M(x_1.d)) & x_1.t = x_1.s + 1 \end{cases}$$

\square

Lemma 4.7. Let $C = (A, f)$ be an MP-LGA then there exist A', f' such that $C' = (A', f')$ is an S-MP-GA and $\forall G \in \mathcal{G}_{\mathcal{B}} \forall v \in V(G) C'(G, v) = C(G, v)$. Furthermore, C' incurs polynomial time and space overhead.

Proof. Unlike in the proof of Lemma 4.5, the emulating function f' is not very concise, hence we define it in pseudo-code style, in Listing 1. Each recurrence of C , where a node simply receives the multiset of its neighbors' features, requires $O(|G|^3)$ recurrences of C' , where a node receives the sum of whatever its neighbors are sending, in order to extract the neighbors' individual features. The idea of such a phase of $O(|G|^3)$ recurrences is as follows:

1. For the first $|G|^2$ recurrences, vertices propagate $\max(\text{own value}, \text{neighbors' messages average})$. The neighbors' average can be computed by dividing the sum of values by the sum of '1' each sender sends a dedicated dimension. Note that if there are two vertices with different values, the lower-value one will necessarily perceive

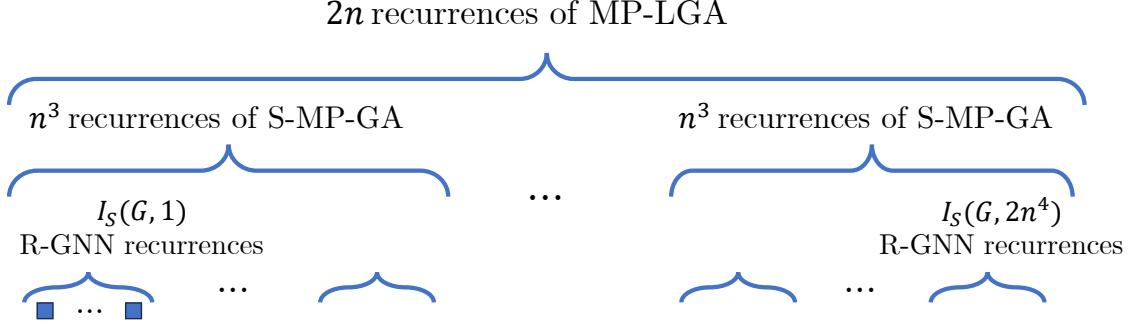


Figure 4: Let $G \in \mathcal{G}_{\mathcal{B}_k}$ and define $n := |G|$. An illustration for reducing the construction of $\text{mpc}_G(v)$ to R-GNN. First it is reduced to $O(n)$ iterations of MP-LGA. Then, each iteration of MP-LGA is reduced to $O(n^3)$ iterations of S-MP-GA. Let that S-MP-GA be S , and let $T_S(G, i)$ be the number of Turing machine steps required to compute the i^{th} iteration of S when operating on G . Then, iteration $i, i \in [2n^4]$ is reduced to $O(I_S(G, i))$ recurrences of an R-GNN, where $I_S(G, i) := T_S(G, i) + (n^3 \log(n) + kn)^2$. The $(n^3 \log(n) + kn)^2$ overhead is for translating the sum of messages from RB to RQ.

an average higher than its own value - even if they are farthest from each other, after at most $|G|$ recurrences. Whenever a vertex perceives a higher average than its own value it temporarily disables itself and propagates the average, until the end of the first $|G|^2$ recurrences. This is implemented in the 'find_max' code. Hence, after $|G|^2$ recurrences, it is guaranteed that the vertices left enabled are exactly those with the maximum value in the whole graph - excluding those already counted for and permanently disabled (see below). The $|G|^2 + 1$ recurrence is used so that each vertex operates according to whether it is one of the max-value vertices. This is implemented in the 'send_if_max' code. Then, in the $|G|^2 + 2$ recurrence each vertex knows that the average it receives is actually the maximum value, and the count is the number of its neighbors with that value, and adds that information to the constructed multiset of neighbors' values. This is implemented in the 'receive_max' code. Then, all max-value vertices permanently disable themselves, all temporarily-disabled vertices re-enable themselves, and another $|G|^2$ recurrences phase starts - to reveal the next maximum-value and counts.

2. Since there are at most $|G|$ distinct values, after $|G|$ iterations of the process above, every vertex has finished constructing the multiset of its neighbors values. All is left to do then is to apply f of the MP-LGA on the (current value and) constructed multiset, and update the vertex's value to be the output of f .

The required message length limit is implied by the definition of the algorithm and the dag-construction description in Section A. \square

Lemma 4.8. *Let $C = (B, h)$ be an S-MP-GA, then there exists an R-GNN $N = (A, F)$ such that $\forall G \in \mathcal{G}_B \ \forall v \in V(G) \ C(G, v) = N(G, v)$. Furthermore, N incurs polynomial time and space overhead.*

Proof. Note that an R-GNN is essentially an extension of a recurrent MLP to the sum-message-passing setting. In

(Siegelmann and Sontag 1992) it is shown that recurrent MLPs are Turing-complete. Assume $C = (A, f)$ and let M be the Turing-machine that computes f , we would like to use the result in (Siegelmann and Sontag 1992) and emulate M using the recurrent MLP in an R-GNN. However, this requires overcoming two significant gaps:

1. An encoding gap. In (Siegelmann and Sontag 1992) the emulation of a Turing machine is done by emulating a two-stack machine where a stack's content is always represented as a value in RQ. Since RQ is not closed under summation, a naive attempt to use the sum of the neighbors' stacks directly - as input to a Turing machine emulation is doomed to fail: The sum may be an invalid input and consequently the output will be wrong. To overcome this, we precede and proceed the Turing-machine-emulation recurrent MLP with recurrent sub-networks that compute translations

$$\text{RB2RQ} : \text{RB} \rightarrow \text{RQ}, \text{RB2RQ} := \tau q \circ \tau b^{-1}$$

$$\text{RQ2RB} : \text{RQ} \rightarrow \text{RB}, \text{RQ2RB} := \tau q \circ \tau b^{-1}$$

The former translates the received sum of messages, and the latter translates the new message to be sent. The depth of these networks is constant and the number of recurrences required for each translation is quadratic for RB2RQ and linear for RQ2RB, in the number of translated bits. Existence of such recurrent MLPs is not trivial - the result in (Siegelmann and Sontag 1992) assumes an input in RQ for a reason.

The messages that are used in our algorithm have a fixed-length, and have two parts which the receiver should be able to read separately. When multiple messages are summed, those two parts can potentially interfere. However, we make sure that the message length is large enough to contain the sums of both parts separately, and assume that they are written in two separate parts of the message, hence it is guaranteed that there will be no interference. Define $L_\delta(n, k) :=$

$$\max(|\delta(\text{mpc}_G(v))| : G \in \mathcal{G}_{\mathcal{B}_k}, |G| = n, v \in V(G))$$

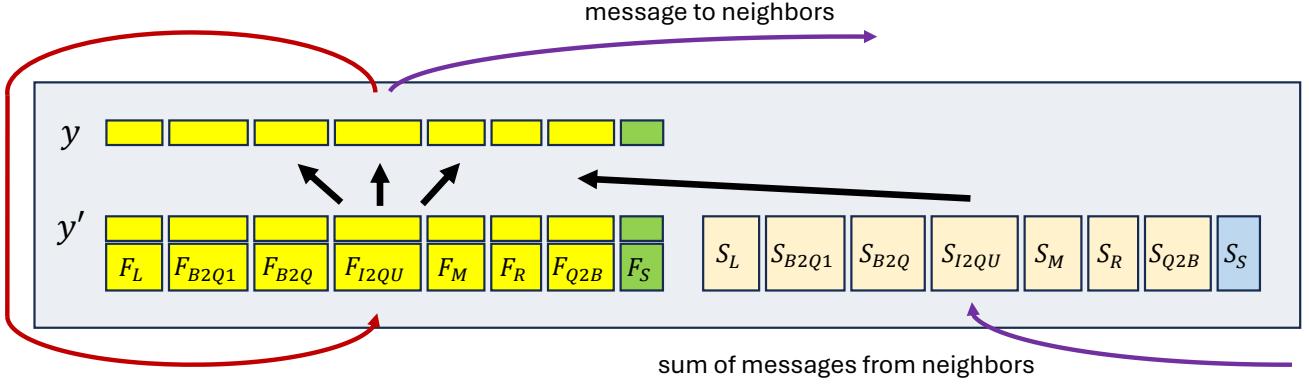


Figure 5: The structure of F . In yellow are all the switched sub-networks that pass control and data from one to another. In green is the synchronization sub-network that runs continuously. In beige and blue are the dimensions that assume the sum of neighbors' sub-networks output in the previous recurrence. Layer y' is the outputs of the individual sub-networks. These, together with the neighbors' sum, are inter-rotated using layers $[y' + 1..y]$, to create the desired interoperability. Finally, the output of layer y is sent to neighbors and also becomes (half) the input for the next recurrence.

the maximum length of the dag encoding of a vertex in a graph of size n with initial features of length k . The fixed message length is $3kn^4$, which is more than the maximum-possibly-required $\log(n(2^{L_\delta(n,k)} + 2^1))$ - the length of the sum of n complete mpc-plus-a-0/1-indicator. The sub-networks are designed to translate those messages. To do that, one of the inputs of the RB2RQ sub-network is the fixed message-length. Note that k is known before constructing the overall network, and n is given as an initial feature of the vertex, however we provide a stronger result by describing a network that receives also k as an input hence supports all initial-feature lengths. We construct a dedicated recurrent sub-network that computes $3kn^4$ at the very beginning of the overall computation, which can then be used by the sub-networks that require this value as one of their inputs.

In addition to the translation of the incoming and outgoing messages, it is required at the beginning of the run to translate the message-length; graph-size; and initial-feature - from an integer and a value in RB to one value in RQ, to be used by the Turing-machine sub-network. This is done in two stages: First, another copy of an RB2RQ translator translates the feature, and then a new kind of sub-network does both the translation of the graph size and the combination of the two RQ values - in a way that they can be separated by the Turing machine emulation.

Finally, As the two final-output dimensions of an R-GNN - a "finished" indicator and the final feature of the vertex - are defined in Definition 4.1 to be a 0/1 value and a value in RB, we construct another translating network that translates those final outputs from RQ, when the whole computation of the R-GNN is about to finish.

2. A synchronization gap. In an S-MP-GA computation, nodes are synchronized by definition: A node's i^{th} received sum-of-messages is the sum of its neighbors i^{th}

sent message. On the other hand, in an emulating R-GNN that is based on (Siegelmann and Sontag 1992), every Turing machine step corresponds to a recurrence and - being a GNN - every recurrence entails a message pass, that is, a potential Send/Receive. A naive attempt of a vertex to start a new computation whenever its emulated Turing machine has finished, hence considering the next message-pass as the sum of its neighbors' current-computations' results, will not work: The (emulated) Turing machines of different nodes with different inputs may require a different number of recurrences to complete their computations. To overcome this, we augment the recurrent MLP described thus far with a recurrent MLP that synchronizes the start of a new computation, across all nodes. That MLP runs for a the same number of recurrences for all nodes - regardless of differences in their inputs, hence, in itself it does not require synchronization.

Overall, the recurrent MLP F consists of 8 recurrent sub-networks:

$F_L, F_{B2Q_1}, F_{I2QU}, F_{B2Q}, F_M, F_R, F_{Q2B}, F_S$. For a graph and vertex $G \in \mathcal{G}_{\mathcal{B}_k}$, $v \in V(G)$, the computation flow is as follows:

- a. F_L computes the length of the messages, $3k|G|^4$, to be used by F_{B2Q} , then passes control to (b).
- b. F_{B2Q_1} translates $Z(G, v)$ from RB to RQ, to be used by F_{I2QU} , then passes control to (c).
- c. F_{I2QU} translates $|G|$ to its unary representation in RQ - base-4 '3' digits, and concatenates it to the base-4 '1' digit followed by the result of (b). That is, $x \mapsto (\sum_{i \in [x]} \frac{3}{4^i}) + \frac{1}{4^{x+1}} + \frac{1}{4^{x+2}} \mathbf{rq}(\mathbf{rb}^{-1}(Z(G, v)))$. The output is to be used by F_M . Then it passes control to (1.1).
- 1.1. F_{B2Q} translates the sum of messages, from RB to RQ, then passes control to (1.2).

- 1.2. F_M computes the Turing machine function, then passes control to (1.3).
- 1.3. F_R Checks whether the overall computation - the function computed by the R-GNN - is finished, and if it is, translates the final output - final feature of the vertex - from RQ to RB. Note that in such case this output is "locked" - it will not change over subsequent recurrences. If the overall computation is not finished, control is passed to (1.4).
- 1.4. F_{Q2B} translates the new message to send, computed by F_M , from RQ to RB, then stops.
2. In parallel to (1), F_S operates continuously. Every $2n$ recurrences it produces a signal whether all nodes' previous computation is finished already n recurrences before. If the signal is positive, the sequence of (1) is restarted.

We proceed to formalize the structure of the whole R-GNN network, that is, the structure of each sub-network and how these sub-networks are put together. We describe them not necessarily in their order of operation - described above - but also in order of similarity in functionality.

We would like the sub-networks $F_L, F_{B2Q_1}, F_{I2QU}, F_{B2Q}, F_M, F_R, F_{Q2B}$ to operate in two modes, 'On' - where they compute what they are supposed to (probably over multiple recurrences) and have the results as the values of certain output neurons, and 'Off' - where the results-outputs are remained unchanged. Switching to 'On' mode should be initiated by an external change to an 'On/Off' switch neuron, at which time certain preconditions must hold for the rest of the state neurons of the sub-network in order to assure that it indeed computes what it is supposed to. Switching to 'Off' mode is part of the sub-network's results when finishing its current computation, which in turn causes another sub-network to switch on, and so on and so forth. The goal is that at any recurrence of the whole R-GNN, only one sub-network, excluding F_S , is actually computing its function. We formalize the concept of a switched network, in the following definition.

Definition B.3. A switched recurrent MLP (switched MLP) $F = ((w_1, b_1), \dots, (w_m, b_m)), D, f)$, of d dimension, is a sequence of rational matrices and bias vectors - similar to a recurrent MLP; A set $D \subseteq \{0, 1\} \times \{0, 1\} \times \mathbb{Q}^{d_I+d_O} \times \mathbb{Q}^{d-d_I-d_O-2}$,

and a function $f : \mathbb{Q}^{d_I} \rightarrow \mathbb{Q}^{d_O}, 2 + d_I + d_O \leq d$. The $d - d_I - d_O$ dimensions are neurons that only remember the state of the computation. The semantics are that the first input dimension is a switch, the second is an indication if the switch has turned off from previous iteration, D is a set of valid initial states, and f is a target function whose input(s) start at the third input dimension, and whose output(s) start after the inputs. More formally:

1. $\forall x \in D \quad x(1) = 0 \Rightarrow (f_F^{(1)}(x)(1) = 0 \wedge f_F^{(1)}(x)(2) = 0 \wedge f_F^{(1)}(x)[2 + d_I + 1, 2 + d_I + d_O] = x[2 + d_I + 1, 2 + d_I + d_O] \wedge f_F^{(1)}(x) \in D)$
- That is, for every valid input such that the first dimension is 0, the MLP remains "switched off": Switch is 'Off',

turned-off indicator is 0, target-function outputs maintain their value, and the overall output is a valid initial state - ready for being switched on.

2. For $x \in D$, define $t_x := \min(t : \forall t' \geq t \quad f_F^{(t')}(x) = f_F^{(t)}(x)) - 1$. Then,

$$\begin{aligned} \forall x \in D \quad &x(1) = 1 \Rightarrow \\ &(f_F^{(t_x)}(x)(1) = 0 \wedge f_F^{(t_x)}(x)(2) = \\ &1 \wedge f_F^{(t_x)}(x)[2 + d_I + 1, 2 + d_I + d_O] = \\ &f(x[3, 2 + d_I]) \wedge f_F^{(t_x)}(x) \in D) \end{aligned}$$

That is, for every valid input such that the first dimension is 1, the MLP is "switched on" and: After enough recurrences, the first output turns 0 - switching off; the switched-off indicator turns on; the function outputs are the value of the target function applied to the function inputs; and the overall output is a valid initial state. Note that by (1) we have that $f^{(t_x+1)}(x)$ is the same as $f^{(t_x)}(x)$ except for the second dimension which is 0 and not 1, because the switch does not turn off in recurrence $t_x + 1$ - it was already off.

Note that while an external input is required to initiate a computation - set on the 'On/Off' switch and initialize the inputs, it is only a preceding stage to the network's operation and not in the scope of its definition. This is in contrast to Definition B.11 where new external inputs are used in every recurrence throughout the computation and hence are part of the networks' specification.

Constructing F_M Remember that we wish the R-GNN to compute the same function as a given S-MP-GA $C = (B, h)$. The first sub-network we describe is the main switched sub-network: At the core, it is a recurrent MLP that emulates the Turing-machine that computes h , which exists by (Siegelmann and Sontag 1992). Then, that MLP is wrapped so to make it a switched recurrent MLP.

Lemma B.4. Let $f : \mathcal{B}^4 \rightarrow \mathcal{B}^4$ be a computable function, and let $g : RQ^4 \rightarrow RQ^4$ such that

$$\forall x \in \mathcal{B}^4 \quad g(\mathbf{rq}(x)) = (\mathbf{rq}(f(x)))$$

Let $S := \{y \in RQ^4 \mid \exists x \in \mathcal{B}^4 \quad y = \mathbf{rq}(x)\}$ be the set of valid inputs to g . Define

$$\forall d' \in \mathbb{N} \quad D_{d'} := \{x \mid x \in \{0, 1\}^2 \times S \times RQ^4 \times \{0, 1\} \times S \times \{0\}^{d'-4}\}$$

Then, for some $d' \in \mathbb{N}$ there exist a $(d' + 11)$ -dimension switched network

$$F = \left(((w_1, b_1), \dots, (w_m, b_m)), D_{d'}, g \right)$$

Proof. Let $d' \in \mathbb{N}$ so there exists a d' -dimension recurrent MLP $F' = (l'_1, \dots, l'_{m'})$ such that

$$\forall x \in S \times \{0\}^{d'-4} \quad \exists k \in \mathbb{N} :$$

$$(\forall t < k \quad F'^{(t)}(x)(5) = 0), \quad F'^{(k)}(x)(5) = 1, \quad F'^{(k)}(x)[1, 4] = g(x)$$

Such F' exists by (Siegelmann and Sontag 1992, Thm 2 and Section 4.3). We define a switched network

$$F = ((l_1, \dots, l_m), D_{d'}, g)$$

of dimension $d := (d' + 11)$ and depth $m := (m' + 5)$, as follows. The idea is to wrap a copy of F' with 2 pre-processing and 3 post-processing layers, and additional dimensions - that relay the information between the pre and post processing parts. Note that later we describe how F_M itself is wrapped when combined into the whole network of the R-GNN.

We denote: The d -dimension input vector to the first layer by x ; the d -dimension input vector to the third layer - the layer in which the copy of F' starts, by x' ; the output vector of layer $m' + 2$, in which the copy of F' ends, by y' ; and the d -dimension output of the last layer, which is the input vector for the next recurrence, by y .

The first 11 dimensions represent F 's switch; switch-turned-off indicator; function inputs; function outputs; and previous-recurrence switch-value. Initially, $x(1) = x(2) = x(11) = 0$ i.e. switch-related values are 0. When F is switched on i.e. $x(1) = 1, x(2) = 0, x(11) = 0$, we want the inputs to become F' inputs and we want to reset the rest of F' dimensions to 0, and so we define

$$\begin{aligned} x'[12, 15] &:= \text{LSig}(x[3, 6] - \text{LSig}(1 - \text{ReLU}(x(1) - x(11)))) + \\ &\quad \text{LSig}(x[12, 15] - \text{LSig}(\text{ReLU}(x(1) - x(11)))) \end{aligned}$$

$$x'[16, d] := \text{LSig}(x[16, d] - \text{LSig}(\text{ReLU}(x(1) - x(11))))$$

We also want the switch-turned-off indicator to always remember the switch state at the beginning of the recurrence, and so we define $x'(2) := x(1)$. Considering that the activation function of our network is ReLU, x' can be computed from x using at most 2 layers. These would be l_1, l_2 .

On the output end, if the input switch was 0, we want all outputs of F' to be zero - so it is ready for a new run. Otherwise: If F' is finished i.e. $y'(16) = 1$ then we want F' function outputs i.e. $y'[12, 15]$ to be the output of F i.e. $y[7, 10]$, its switch to turn off, and the switch-turned-off indicator to be 1. If F' is not finished then we want to pass all outputs of F' as they are to next recurrence. The above logic can be implemented as follows:

$$y(1) = \text{ReLU}(y'(1) - y'(16)) \text{ (if } F' \text{ finished then turn off)}$$

$$y(2) = \text{ReLU}(y'(2) - \text{ReLU}(y'(1) - y'(16))) \text{ (if } F' \text{ just turned off)}$$

$$y(11) = \text{ReLU}(y'(1) - y'(16))$$

$$\begin{aligned} y[7, 10] &= \text{ReLU}\left(y'[12, 15] - \text{ReLU}(1 - y'(16))\right) + \\ &\quad \text{ReLU}\left(y'[7, 10] - y'(16)\right) \end{aligned}$$

Considering that the activation function of our network is ReLU, y can be computed from y' using at most 3 layers. These would be $w_{2+m'+1}, w_{2+m'+2}, w_{2+m'+3}$. Layers $[3..m'+2]$ are essentially 2 separate stack-of-layers side-by-side, as follows:

1. The first stack is F 's (switch, switch-turned-off indicator, function inputs, function outputs, previous switch), in dimensions $[1..11]$, each simply passing on from layer to layer.

2. The second column is F' as is, set in layers $[3..m'+2]$. □

For F_M , we did not explicitly describe the core network that implements the required functionality. Instead, we deduced its existence, from the result in (Siegelmann and Sontag 1992). However, the mode of operation of a network ala (Siegelmann and Sontag 1992) does not match the mode of operation of an R-GNN. As the combined functionality of $F_L, F_{B2O_1}, F_{I2QU}, F_{B2Q}, F_M, F_R, F_{Q2B}$ is meant to bridge that gap, the descriptions of these sub-networks cannot themselves be based on (Siegelmann and Sontag 1992) since this would just move the gap to a different location in the algorithm. Hence, we describe each of these sub-networks explicitly: We provide a pseudo-code that implements the required functionality when ran recurrently, and meets a certain specification which implies implementability by an MLP. The specification consists of a structure and an instruction-set, which together define an MLP implementation, as follows.

Programming MLPs

Definition B.5. An MLP Code having d state-variables i.e. variables that should maintain their value from the end of one recurrence to the beginning of the next, has the following properties which together imply a d -dimension MLP implementation.

- The code has two sections:

An 'Initialize' section. There, the d state-variables, which correspond to the I/O neurons of the MLP, are defined. Their names are prefixed by 's_', and their initial values - before the R-GNN run starts - are defined as well. These must be non-negative.

A 'RecurrentOp' section. There, is the code that defines a single recurrence of an implementing MLP. It consists solely of statements that conform to one of the following definitions of an expression. The invariant of these definitions is that for every expression there exists an MLP sub-structure whose inputs are the statement's operands and whose output is the value of the statement.

- The state-variables are expressions. They correspond to the input neurons, and to subsequent dedicated neurons in each layer. That is, for each state-variable there is a corresponding neuron in each layer.
- For the sake of code-readability, intermediate variables may also be used, and are considered expressions. Declared using the keyword 'var', they correspond to a specific neuron whose value is defined by an assignment to the variable.
- Let x_1, \dots, x_n be expressions, let $w_1, \dots, w_n \in \mathbb{Q}, b_1, \dots, b_n \in \mathbb{Q}$, and define $x := \sum_{i \in [n]} w_i x_i + b_i$, then:
 1. Both $\text{ReLU}(x), \text{LSig}(x)$ are expressions, regardless of the sign of x . Note that LSig is expressible by combining two ReLUs.
 2. If $x \geq 0$ then x is an expression.

- Let v, x be a variable (expression) and an expression, then $v = x$ is an expression. Note that such assignment is implemented by feeding the neuron that represents x into a next-layer neuron that represents v .
- Let v be a variable, and let $0 \leq x \leq 1, s \in \{0, 1\}$ be expressions representing a change and a state. Then, the conditional addition

$$v.\text{increase_if}(s, x) := \begin{cases} v + x & s = 1 \\ v & s = 0 \end{cases}$$

is an expression, and if $v - x > 0$ then also the conditional subtraction,

$$v.\text{decrease_if}(s, x) := \begin{cases} v - x & s = 1 \\ v & s = 0 \end{cases}$$

is an expression. Note that due to the restrictions $0 \leq x \leq 1, s \in \{0, 1\}$, $v.\text{increase_if}(s, x) \equiv v + \text{ReLU}(x - (1 - s))$. Similar argument holds for $v.\text{decrease_if}(s, x)$.

- Let $0 \leq v \leq 1$ be a variable, let $a \in \mathbb{N}$ be a constant, and let $s \in \{0, 1\}$ be an expression. Then, the conditional division

$$v.\text{diva_if}(s) := \begin{cases} \frac{v}{a} & s = 1 \\ v & s = 0 \end{cases}$$

is an expression. Note that due to the restrictions on the arguments, $v.\text{diva_if}(s) \equiv v - \text{ReLU}((1 - \frac{1}{a})v - (1 - s))$, which is expressible by a neuron.

- Let $0 \leq v \leq 1$ be a variable, let $0 \leq x \leq 1$ be an expression, and let $s \in \{0, 1\}$ be another expression. Then, the conditional assignment

$$v.\text{set_if}(s, x) := \begin{cases} x & s = 1 \\ v & s = 0 \end{cases}$$

is an expression. Note that due to the restrictions on the arguments, $v.\text{set_if}(s, x) \equiv v.\text{decrease_if}(s, v) + \text{ReLU}(x - (1 - s))$, which is expressible by 2 neurons in one layer and another neuron in the next layer.

Constructing F_L

Lemma B.6. Let $a \in \mathbb{Q}$, then there exist $d \in \mathbb{N}$ for which there is a d -dimension switched network

$$F = ((l_1, \dots, l_m), \{(1, 0), (0, 1), (0, 0)\} \times \mathbb{N}^2 \times \mathbb{Q}^8, (x_1, x_2) \rightarrow ax_1x_2^3)$$

That is, there exists a switched network that for every $a \in \mathbb{Q}, x_1, x_2 \in \mathbb{N}$ computes $ax_1x_2^3$.

Proof. Following the code in Listing 2, which conforms to Definition B.5, it is not difficult to verify that it defines a 12-dimension switched network

$$F = \left(((w_1, b_1), \dots, (w_m, b_m)), \{(1, 0), (0, 1), (0, 0)\} \times \mathbb{N}^2 \times \mathbb{Q}^8, (x_1, x_2) \rightarrow ax_1x_2^3 \right)$$

for some rational matrices and biases (w_i, b_i) . \square

Constructing F_{B2Q_1} The sub-network F_{B2Q_1} is identical to F_{B2Q} which is described later. The difference is that F_{B2Q_1} is used only once, in the initialization stage of the whole $R - GNN$ run. There, it converts the input feature, provided by the user in RB encoding, to RQ encoding, which is then used by F_{I2QU} .

Constructing F_{I2QU}

Lemma B.7. Define $D := \{(1, 0), (0, 1), (0, 0)\} \times \mathbb{N} \times RQ \times \mathbb{Q}_{[0,1]} \times \{0\}$, and define $\forall x \in \mathbb{N} \ \forall y \in RQ \ g(x, y) := (\frac{1}{4^{x+1}} + \frac{y}{4^{x+2}} + \sum_{i \in [x]} \frac{3}{4^i})$. Then, there exists a switched network

$$F = ((l_1, \dots, l_m), D, g)$$

That is, there exists a switched network that given $x \in \mathbb{N}$ and $y \in RQ$ outputs the translation of x to Unary Rational Quaternary concatenated with the digit '1' in RQ concatenated to y .

Proof. Following the code in Listing 3, which conforms to Definition B.5, it is not difficult to verify that it defines the required switched network. \square

Constructing F_{B2Q}

Lemma B.8. Define

$$\forall L \in \mathbb{N} \ D_L := \{(1, 0), (0, 1), (0, 0)\} \times \{L\} \times \{x : x = \text{rb}(y), y \in \mathcal{B}_L\} \times \mathbb{Q}_{[0,1]} \times \mathbb{Q}^{12}$$

Then, there exists a recurrent MLP $F = (l_1, \dots, l_m)$ such that $\forall L \in \mathbb{N}$ it holds that $S_L = (F, D_L, \text{rq} \circ \text{rb}^{-1})$ is a switched network. That is, there exists a recurrent MLP such that given a message length $L \in \mathbb{N}$ and a rational binary encoding of a message $x \in \mathcal{B}_L$, outputs the message's rational quaternary encoding.

Proof. Following the code in Listing 4, which conforms to Definition B.5, it is not difficult to verify that it defines the required switched network. \square

Constructing F_{Q2B} Unlike F_{B2Q} , F_{Q2B} does not have the message length as input, since the RQ encoding allows us to identify, without receiving the length explicitly, when we have read all digits and the message is empty: The value of an empty message is 0 while the value of any non-empty message is at least $\frac{1}{4}$.

Lemma B.9. Define $D := \{(1, 0), (0, 1), (0, 0)\} \times RQ \times \mathbb{Q}_{[0,1]} \times \{0\} \times \mathbb{Q}_{[0,1]}$, then there exists a switched network

$$F = ((l_1, \dots, l_m), D, \text{rb} \circ \text{rq}^{-1})$$

That is, there exists a switched network that for every $x \in \mathcal{B}$ translates its rational quaternary encoding to its rational binary encoding.

Proof. Following the code in Listing 5, which conforms to Definition B.5, it is not difficult to verify that it defines the required switched network. \square

Constructing F_R

Lemma B.10. Define $D := \{(1, 0), (0, 1), (0, 0)\} \times RQ^2 \times \mathbb{Q}_{[0,1]}^2 \times \{0\} \times \mathbb{Q}_{[0,1]}\}$, then there exists a switched network

$$F = ((l_1, \dots, l_m), D, (1, \text{rb} \circ \text{rq}^{-1}))$$

That is, there exists a switched network that for every $x \in \mathcal{B}$ outputs the indication 1 and the translation of x from its rational quaternary encoding to its rational binary encoding.

Proof. Following the code in Listing 6, which conforms to Definition B.5, it is not difficult to verify that it defines the required switched network. \square

The last sub-network we describe, F_S , is not a switched network. It is 'on' in every recurrence, in parallel to whatever switched network is 'on' in that recurrence. Also, part of its input dimensions are external i.e. they do not assume their output-layer value from the previous recurrence, rather, they are set externally at the beginning of each recurrence. We formally define this kind of network as follows.

Definition B.11. A d -dimension m -depth recurrent MLP with external inputs

$F = ((l_1, \dots, l_m), z_1, \dots, z_L)$ is a variation of a d -dimension recurrent MLP, such that the last L dimensions are external input sequences $\{z_i^{(t)}\}_{i \in [L], t \in \mathbb{N}}$. That is,

$$\begin{aligned} f_F^{(0)}(x) &:= x \in \mathbb{Q}^d, \quad \forall t > 0 \quad f_F^{(t)}(x) := \\ f_F\left(f_F^{(t-1)}(x)[1, d-L], (z_1(t), \dots, z_L(t))\right) \end{aligned}$$

Constructing F_S The sub-network F_S implements a mechanism that synchronizes the nodes' Turing-machine emulations. The idea is that at specific recurrences, all nodes record and broadcast their status. Then, for the next $|G|$ recurrences they consider their own recording and the messages they receive, and broadcast whether there is an indication of a non-finished node. This means that if and only if there is such node then all nodes will be aware of it at the end of those $|G|$ recurrences - because it takes at most $|G|$ recurrences for the information to propagate. In other words, all nodes have the same global status-snapshot at the end of those $|G|$ recurrences.

If according to the snapshot all nodes' Turing machines are finished, then all nodes start their new computation in the next recurrence - exactly at the same time. In addition, we construct the R-GNN such that when a node's Turing machine is finished the node continues to send the same result-message in subsequent recurrences until starting a new computation. The above scheme assures that all nodes' Turing machines start computation $i + 1$ - emulating iteration $i + 1$ of the emulated S-MP-GA algorithm - with input which contains the sum of their neighbors' i^{th} -computation results.

Due to technical limitations of computing with MLP, F_S runs in cycles of $2|G|$ recurrences instead of $|G|$: The first $|G|$ recurrences are used to reset its state, and the second $|G|$ recurrences implement the scheme above. It has three external inputs: The graph size; the sum of neighbors' "some

node not finished yet at the beginning of the cycle" indicator; a 'finished' indicator for the node's current computation, and two outputs: "should start new computation" - used by F_{B2Q} ; and "some node not finished yet at the beginning of the cycle" indicator - to be sent to neighbors. The first output is 0 throughout the cycle except for the last recurrence where it is 1 in case all nodes are finished, and the second output is 0 throughout the first half of the cycle - where it is basically disabled - and becomes 1 in the second half in case an indication of a non-finished node was received.

The following Lemma formalizes the described behavior.

Lemma B.12. Let $n \in \mathbb{N}$, then there exist a d -dimension recurrent with external inputs MLP

$$F = ((l_1, \dots, l_m), z_1, z_2, z_3) \text{ such that:}$$

$$f_F^{(t)}(1) := \begin{cases} 0 & t \bmod 2n > 0 \\ 1 - f_F^{(t-1)}(2) & t \bmod 2n = 0 \end{cases}$$

$$f_F^{(t)}(2) := \begin{cases} 0 & t \bmod 2n \leq n \\ \min(1, \sum_{i=t-(t \bmod n)}^t (z_2^{(i)} + z_3^{(i)})) & t \bmod 2n > n \end{cases}$$

Proof. Following the code in Listing 7, which conforms to Definition B.5, it is not difficult to verify that it defines the required recurrent network with external inputs. \square

Constructing The R-GNN $N = (A, F)$ We are now ready to put together all the sub-networks above into one recurrent MLP F . Their I/O dimension, initial values - corresponding to their part in the A vector, and individual functionality, are already defined in their dedicated subsections above. Here we describe how they are connected to one another to form the whole recurrent network F . We refer to them by their names in the subsections describing their construction. Please refer to Figure 5 for an illustration of the construction.

Remark 1. Note that the positions of the inputs to the various sub-networks, as implied by the sub-networks' definitions and the construction below, do not match their positions according to the definition of $N^{(0)}(G, v)$ in Definition 4.1. The reason for the permutation is so both the definition of $N^{(0)}(G, v)$ and the description of the construction are easier to follow. Reversing the permutation e.g. by adding a prelude and postlude layers is trivial and for purpose of focus we do not include it.

For a d -dimension m -depth sub-network F_H we define $d(F_H) := d$; $m(F_H) := m$. Note that when referring to the dimensions of a sub-network we mean its I/O dimensions. Whatever are the dimensions and arguments of its hidden layers, they do not make a difference to our construction of F . Assume that F_H is embedded in F such that dimensions $a..(a + d(F_H) - 1)$ in the input layer of F are the input layer of F_H . We define $E(F_H) := (a + d(F_H) - 1)$ the last dimension of the part of F_H in F .

Define $m_{\max} := \max(m(F_H) \mid H \text{ is a sub-network})$ the maximum depth over all sub-networks. We assume all sub-networks of lower depth are extended to depth m_{\max} by simply passing on their output layer, hence we can refer to layer

m_{\max} of any sub-network. We denote by $y'_{F_H}(i)$ the i^{th} dimension in the output layer of the F_H part of F , that is, the $a + i - 1$ neuron in layer m_{\max} of F . To implement the interoperability between the sub-networks, we add additional layers after y' , up to the final output layer which we denote by y . The interconnections are defined by describing certain dimensions in y as functions of (also) dimensions in other sub-networks' outputs and the aggregation values - in y' . Dimensions in y for which we do not define functions, simply assume their values in y' as is. We denote by $y_{F_H}(i)$ the $a + i - 1$ neuron in the last layer of F - whose output in one recurrence is the value of the $a + i - 1$ input neuron in the following recurrence. Note that y, y' are not to be confused with the y, y' layers mentioned in the individual description of F_M .

Let H be the unique name of one of the sub-networks e.g. $H = B2Q$. We denote by S_H the dimensions of F that are the sum of the neighbors' F_H . For example, $S_M(1)$ is the sum, over all neighbors, of the first dimension of F_M 's output in the previous recurrence, and is in dimension $E(F_S) + E(F_{B2Q}) + 1$ in F . As described later, the sum values are used as inputs, only in the extra layers of F - after the y' layer. That is, they simply pass on from the input layer until the last layer, and some of their neurons are inputs to other parts of F in the extra layers after y' .

And so, the whole flow is established in which at every recurrence of F only specific sub-networks are active, and eventually their outputs become inputs for other sub-networks which in turn become the active ones etc.

Subnetwork F_L , dimensions $[1..d(F_L)]$

Subnetwork F_{B2Q_1} , dimensions $[d(F_L) + 1..d(F_L) + d(F_{B2Q_1})]$

$y_{F_{B2Q_1}}(1) := \text{ReLU}(y'_{F_{B2Q_1}}(1) + y'_{F_L}(2))$ i.e. F_{B2Q_1} should start after F_L is finished.

$y_{F_{B2Q_1}}(3) := \text{ReLU}(y'_{F_L}(3) - (1 - y'_{F_L}(2))) + \text{ReLU}(y'_{F_{Q2B_1}}(3) - (y'_{F_L}(2)))$ i.e. F_{Q2B_1} feature length input is set to be the dimension with that data used also by F_L , for the beginning of a new computation, otherwise it maintains its value.

Subnetwork F_{I2UQ} , dimensions $[E(F_{B2Q_1}) + 1..E(F_{B2Q_1}) + d(F_{I2UQ})]$

$y_{F_{I2UQ}}(1) := \text{ReLU}(y'_{F_{I2UQ}}(1) + y'_{F_{B2Q_1}}(2))$ i.e. F_{I2UQ} should start after F_{B2Q_1} is finished.

$y_{F_{I2UQ}}(3) := \text{ReLU}(y'_{F_L}(4) - (1 - y'_{F_{B2Q_1}}(2))) + \text{ReLU}(y'_{F_{I2UQ}}(3) - (y'_{F_{B2Q_1}}(2)))$ i.e. F_{I2UQ} graph-size input is set to be the dimension with that data used also by F_L , for the beginning of a new computation, otherwise it maintains its value.

$y_{F_{I2UQ}}(4) := \text{ReLU}(y'_{F_{B2Q_1}}(5) - (1 - y'_{F_{B2Q_1}}(2))) + \text{ReLU}(y'_{F_{I2UQ}}(4) - (y'_{F_{B2Q_1}}(2)))$ i.e. F_{I2UQ} num-in-RQ input is set to be the output of F_{B2Q_1} , for the beginning of a new computation, otherwise it maintains its value.

Subnetwork F_{B2Q} , dimensions $[E(F_{I2UQ}) + 1..E(F_{I2UQ}) + d(F_{B2Q})]$

$y_{F_{B2Q}}(1) := \text{ReLU}(y'_{F_{B2Q}}(1) + y'_{F_S}(1) - y'_{F_{I2UQ}}(1))$ i.e. F_{B2Q} should start only after F_{I2UQ} has switched off, and from then on it should be on either if it is in the middle of a computation or if it received a signal from F_S that it can start again. Note that a signal from F_S means that $F_M; F_{Q2B}$ are off.

$y_{F_{B2Q}}(3) := y'_{F_L}(5)$ i.e. F_{B2Q} should take $L_{n,k}$ from the first output of F_L .

$y_{F_{B2Q}}(4) := \text{ReLU}(y'_{S_M}(8) - y'_{F_S}(1) - y'_{F_{I2UQ}}(1))$ i.e. F_{B2Q} should take new input from the sum of neighbors (at position: second output of F_M) when it is signaled to start again.

Subnetwork F_M , dimensions $[E(F_{B2Q}) + 1..E(F_{B2Q}) + d(F_M)]$

$y_{F_M}(1) := \text{ReLU}(y'_{F_M}(1) + y'_{F_{B2Q}}(2))$ i.e. F_M should turn on after F_{B2Q} turns off.

$y_{F_M}(3) := \text{ReLU}(y'_{F_M}(7) - (1 - y'_{F_{B2Q}}(2))) + \text{ReLU}(y'_{F_M}(3) - (y'_{F_{B2Q}}(2)))$ i.e. F_M first input is set to be the first output for the beginning of a new computation, otherwise it maintains its value.

$y_{F_M}(4) := \text{ReLU}(y'_{F_{B2Q}}(5) - (1 - y'_{F_{B2Q}}(2))) + \text{ReLU}(y'_{F_M}(4) - (y'_{F_{B2Q}}(2)))$ i.e. F_M second input is set to be the output of F_{B2Q} for the beginning of a new computation, otherwise it maintains its value.

$y_{F_M}(7) := \text{ReLU}(y'_{F_{I2UQ}}(5) - (1 - y'_{F_{I2UQ}}(2))) + \text{ReLU}(y'_{F_M}(7) - (y'_{F_{I2UQ}}(2)))$ i.e. F_M first output gets the result of F_{I2UQ} and maintains it as long as there is no computation, so when the first computation starts $y_{F_M}(3)$ will read the the result of F_{I2UQ} .

Subnetwork F_R , dimensions $[E(F_M) + 1..E(F_M) + d(F_R)]$

$y_{F_R}(1) := \text{ReLU}(y'_{F_R}(1) + y'_{F_M}(2))$ i.e. F_R should turn on after F_M turns off.

$y_{F_R}(3) := \text{ReLU}(y'_{F_M}(10) - (1 - y'_{F_M}(2))) + \text{ReLU}(y'_{F_R}(3) - (y'_{F_M}(2)))$ i.e. F_R 'allFinished' input is set to be the last output of F_M upon start.

$y_{F_R}(4) := \text{ReLU}(y'_{F_M}(9) - (1 - y'_{F_M}(2))) + \text{ReLU}(y'_{F_R}(4) - (y'_{F_M}(2)))$ i.e. F_R 'final feature' input is set to be the one before last output of F_M upon start.

Subnetwork F_{Q2B} , dimensions $[E(F_R) + 1..E(F_R) + d(F_{Q2B})]$

$y_{F_{Q2B}}(1) := \text{ReLU}(y'_{F_{Q2B}}(1) + y'_{F_R}(2) - y'_{F_R}(5))$ i.e. F_{Q2B} should turn on after F_R turns off but not if overall computation of the R-GNN is finished.

$y_{F_{Q2B}}(3) := \text{ReLU}(y'_{F_M}(8) - (1 - y'_{F_R}(2))) + \text{ReLU}(y'_{F_{Q2B}}(3) - (y'_{F_R}(2)))$ i.e. F_{Q2B} input is set to be the second output of F_M for the beginning of a new computation, otherwise it maintains its value.

Subnetwork F_S , dimensions $[E(F_{Q2B}) + 1..E(F_{Q2B}) + d(F_S)]$

$y_{F_S}(d(F_S) - 2) := y'_{F_L}(2)$ i.e. F_S first input is the graph

size, which is found constantly in the second dimension of F_L .

$y_{F_S}(d(F_S) - 1) := \left(1 - (y'_{F_{B2Q}}(1) + y'_{F_M}(1) + y'_{F_{Q2B}}(1) + y'_{F_{B2Q}}(2) + y'_{F_M}(2) + y'_{F_{Q2B}}(2) + y'_{F_R}(2)) \right)$ i.e. F_S second input should be 1 if the whole computation process is currently off i.e. most recent computation is finished. The reason for sampling both the "switch on" and "switch just turned off" dimensions is to avoid getting into the analysis of corner-case-timing cases.

$y_{F_S}(d(F_S)) := y'_{S_S}(2)$ i.e. F_S third input should be the sum of the neighbors' F_S indication if it has received a "not-finished" signal. \square

Listing 1: Emulate MP-LGA

```

1
2 Initialize: // implementation of  $C^{(0)}(G, v)$ 
3     dim1.graph_size = graph_size
4     dim1.feature = feature
5     dim1.disabled = false
6     dim1.tmp_disabled = false
7     dim1.neighbors_values_and_counts = {}
8     dim1.inner_loop_counter = 0
9     dim1.outer_loop_counter = 0
10    dim1.receive_max = 0
11    dim1.MP_GC_dim1 = MP_GC_init_dim1
12    dim1.MP_GC_iteration_count = 0
13    dim2.count = 1
14    dim2.value = feature
15    dim3 = 0
16    dim4 = 0
17
18 run(prev_dim1, neighbors.dim2.sum, prev_dim3){
19     output_dim1 = prev_dim1.copy() // start with a copy, then set what needs to be updated
20     output_dim3 = prev_dim3 // used to indicate finishing the overall computation, start with same value as previous
21     output_dim4 = prev_dim4 // used to hold the final value when the computation is finished
22     if(prev_dim1.MP_GC_iteration_count == prev_dim1.graph_size+1){// finished whole computation, output.dim2.value should have
23         ↪ the final output
24         output_dim3 = 1
25         output_dim4 = prev_dim1.feature
26     }
27     else if(prev_dim1.outer_loop_counter == prev_dim1.graph_size){// finished collecting multiset of neighbors' features, run
28         ↪ MP_GC_func
29         (MP_GC_output_dim1, MP_GC_output_dim2) = MP_GC_func(prev_dim1.MP_GC_dim1, prev_dim1.
30             ↪ neighbors_values_and_counts)
31         output_dim1.MP_GC_dim1 = MP_GC_output_dim1
32         output_dim1.feature = MP_GC_output_dim2
33         output_dim1.MP_GC_iteration_count = prev_dim1.MP_GC_iteration_count + 1
34         output_dim1.neighbors_values_and_counts = {}
35         output_dim1.inner_loop_counter = 0
36         output_dim1.outer_loop_counter = 0
37     }
38     else if(prev_dim1.outer_loop_counter < prev_dim1.graph_size){// still collecting
39         if(prev_dim1.inner_loop_counter==prev_dim1.graph_size^2){ // finished isolating max among uncollected
40             send_if_max(prev_dim1, output_dim1, output_dim2)
41             output_dim1.receive_max = 1 // next stage is to read neighbors that sent max
42             output_dim1.inner_loop_counter = 0
43         }
44         else if(prev_dim1.receive_max == 1){
45             receive_max(prev_dim1, neighbors.dim2.sum, output_dim1, output_dim2)
46             output_dim1.receive_max = 0
47             output_dim1.outer_loop_counter +=1
48         }
49         else{
50             find_max(prev_dim1, neighbors.dim2.sum, output_dim1, output_dim2)
51             output_dim1.inner_loop_counter +=1
52         }
53     }
54     find_max(prev_dim1, neighbors.dim2.sum, output_dim1, output_dim2){
55         neighbors_avg_value = neighbors.dim2.sum.value / neighbors.dim2.sum.count
56         output.dim=2 = 1
57         // if the vertex is disabled (or tmpDisabled) or its initial feature is lower than the observed value, then propagate the observed value
58         output.dim2.value = neighbors_avg_value
59         if(prev_dim1.feature < neighbors_avg_value)
60             {
61                 // vertex value is lower, then tmpDisable vertex so eventually the only non-tmpDisabled vertices will be those with maximum

```

```

61           ↪ value among the enabled vertices in the graph.
62           output.dim1.tmp_disabled = true
63     }
64   else if (!prev.dim1.disabled && !prev.dim1.tmp_disabled)
65   {
66     // if the vertex is enabled and its initial feature higher than the observed value then propagate its value
67     output.dim2.value = prev.dim1.feature
68   }
69 }
70
71 send_if_max(prev.dim1, output.dim1, output.dim2){
72   if (!prev.dim1.disabled && !prev.dim1.tmp_disabled)
73   {
74     // vertex is one of those with max value among the enabled, send its value to its neighbors, and disable it
75     output.dim1.disabled = true
76     output.dim2.count = 1
77     output.dim2.value = prev.dim1.feature
78   }
79 else
80   {
81     // vertex is either disabled because it already sent its (relatively high) value, or it is
82     // tmpDisabled because of its relatively low value. Then, signal that its shouldn't be counted
83     output.dim2.count = 0
84     output.dim2.value = 0
85   }
86 }
87
88 receive_max(prev.dim1, neighbors.dim2.sum, output.dim1, output.dim2){
89   neighbors.avg_value = neighbors.dim2.sum.value / neighbors.dim2.sum.count
90   // we assume that at this point the vertices that sent a non-zero value, and '1' dim2.count, all share the same value – the maximum
91   // ↪ value among non-disabled vertices in the graph. Hence, their average is that maximum value.
92   if (neighbors.avg_value > 0)
93     // otherwise the vertex has no neighbors with the max value, since we assume non-zero initial values
94   {
95     output.dim1.neighbors_values_and_counts.add(neighbors.dim2.sum.count, neighbors.avg_value)
96   }
97   if (!prev.dim1.disabled)
98     // vertex still hasn't got to be a max non-disabled value, hence it continues to try – until all higher values will be recorded.
99   {
100     output.dim1.tmp_disabled = false
101     output.dim2.count = 1
102     output.dim2.value = prev.dim1.initial_feature
103   }
104 else
105   {
106     output.dim2.count = 0
107     output.dim2.value = 0
108   }
109 }
```

Listing 2: Implement Message Length Computation

```

1 // Computes 3·s.initFeatLen·s.graphSize4
2
3 Initialize:
4   [1] s_switchOn = 1
5   [2] s_switchTurnedOff = 0
6   [3] s_initFeatLen = 0
7   [4] s_graphSize = graphSize
8   [5] s_result = 0
9   [6] s_counter1 = 0
10  [7] s_counter2 = 0
11  [8] s_counter3 = 0
12  [9] s_counter4 = 0
13  [10] s_stateReset1 = 0
14  [11] s_stateReset2 = 0
15
```

```

16 [12] s_stateReset3 = 0
17 [13] s_stateAddTo1 = 0
18
19 RecurrentOp:
20     var prevSwitchVal = s_switchOn
21     var edgeCaseOne = Lsig(1 - (s_input1 - 1)) // s_input1 = 1
22     // we add 3 times because we want to multiply by 3
23     s_result.increase_if(LSig(s_switchOn - s_stateReset1 - s_stateReset2 - s_stateReset3 ), 1)
24     s_result.increase_if(LSig(s_switchOn - s_stateReset1 - s_stateReset2 - s_stateReset3 ), 1)
25     s_result.increase_if(LSig(s_switchOn - s_stateReset1 - s_stateReset2 - s_stateReset3 ), 1)
26
27     var goodOne = LSig(LSig(s_graphSize - s_counter1) - (1 - s_stateAddTo1) - edgeCaseOne) // we want to add and we can
28     s_stateReset1 = LSig(s_stateReset1 - LSig(1 - s_counter1)) // maintain
29     s_stateReset1 = LSig(s_stateReset1 + LSig(s_stateAddTo1 - LSig(s_graphSize - 1 - s_counter1))) // turn on: we want to add counter
30     ↪ reached limit, reset counter
31     s_stateReset1 = LSig(s_stateReset1 - edgeCaseOne)
32     s_stateAddTo1 = \lsig(s_stateAddTo1 - s_stateReset1 - edgeCaseOne);
33     s_counter1.decrease_if(s_stateReset1, 1) // resetting
34     s_counter1.increase_if(goodOne, 1)
35
36     var addTo2 = LSig(LSig(s_stateReset1 - LSig(s_counter1))) // reset1 finished
37     var goodTwo = LSig(LSig(s_graphSize - s_counter2) - (1 - addTo2)) // we want to add and we can
38     s_stateReset2 = LSig(s_stateReset2 - LSig(1 - (s_counter2))) // maintain
39     s_stateReset2 = \lsig(s_stateReset2 + \lsig(addTo2 - \lsig(s_graphSize - 1 - s_counter2))); // turn on: we want to add but cannot
40     s_counter2.decrease_if(s_stateReset2, 1)
41     s_counter2.increase_if(goodTwo, 1)
42
43     var addTo3 = LSig(LSig(s_stateReset2 - LSig(s_counter2))) // reset2 finished
44     var goodThree = LSig(LSig(s_graphSize - 1 - s_counter3) - (1 - addTo3)) // we want to add and we can
45     s_stateReset3 = LSig(s_stateReset3 - LSig(1 - (s_counter3))) // maintain
46     s_stateReset3 = LSig(s_stateReset3 + LSig(addTo3 - LSig(s_graphSize - 1 - s_counter3))) // turn on: we want
47     s_counter3.decrease_if(s_stateReset3, 1)
48     s_counter3.increase_if(goodThree, 1)
49
50     var addTo4 = LSig(LSig(s_stateReset3 - LSig(s_counter3))) // reset3 finished
51     var goodFour = LSig(LSig(s_graphSize - 1 - s_counter4) - (1 - addTo4)) // we want to add and we can
52     s_stateReset4 = LSig(s_stateReset4 - LSig(1 - (s_counter4))) // maintain
53     s_stateReset4 = LSig(s_stateReset4 + LSig(addTo4 - LSig(s_graphSize - 1 - s_counter4))) // turn on: we want
54     s_counter4.decrease_if(s_stateReset3, 1)
55     s_counter4.increase_if(goodThree, 1)
56
57     var addTo5 = \lsig(\lsig(s_stateReset4 - \lsig(s_counter4))); // reset3 finished
58     var goodFive = LSig(LSig(s_initFeatLen - 1 - s_counter5) - (1 - addTo5)) // we want to add and we can
59     s_counter5.increase_if(goodFive, 1)
60
61     s_switchOn = 1 - LSig(LSig(s_counter5 - s_initFeatLen + 2) + addTo5 + LSig(1 - goodFive) - 2)
62     s_stateAddTo1 = LSig(1 - s_stateReset1 - s_stateReset2 - s_stateReset3 - s_stateReset4 - (1 - s_switchOn))
63
64     s_switchTurnedOff = ReLU(prevSwitchVal - s_switchOn)

```

Listing 3: Implement I2QU Translation + Concatenation To Other

```

1 Initialize:
2     [1] s_switchOn = 0
3     [2] s_switchTurnedOff = 0
4     [3] s_numberLeftToProcess = 0
5     [4] s_otherInRQ = 0
6     [5] s_result = 0
7     [6] s_stateAddToNumber = 0
8
9 RecurrentOp:
10    var prevSwitchVal = s_stateAddToNumber
11    s_stateAddToNumber = LSig(s_numberLeftToProcess + s_switchOn - 1)
12    s_switchOn = s_stateAddToNumber
13

```

```

14 s.switchTurnedOff = ReLU(prevSwitchVal - s.switchOn)
15 var switchTurnedOn = ReLU(s.switchOn - prevSwitchVal)
16
17 // init procedure, when switch turns on
18 s.result.set_if(switchTurnedOn, s.otherInRQ)
19 s.result.div4_if(switchTurnedOn) // together with next line: insert a separating "0" i.e. 1/4 in RQ
20 s.result.increase_if(switchTurnedOn, 1/4.0)
21
22 // operation in add to number state
23 s.result.div4_if(s.stateAddToNumber)
24 s.result.increase_if(s.stateAddToNumber, 3/4.0)
25 s.numberLeftToProcess.decrease_if(s.stateAddToNumber, 1)

```

Listing 4: Implement B2Q Translation

```

1 /* The idea of the algorithm in general lines is as follows:
2 Initially,  $x = \sum_{i \in [m]} \frac{a_i}{2^i}$  where  $x = s.number\_in\_process$ ,  $m = s.messageBitLength$ .
3 All relevant variables are reset to their starting values in the s.stateInit stage. Then,
4 For i=1..m
5 Assume we are left with  $x = \sum_{j \in [i..m]} \frac{a_j}{2^j}$ . The s.stateReduce stage implements:
6  $x = \text{ReLU}(\sum_{j \in [i+1..m]} \frac{1}{2^j})$  // at that point  $a_i = 1 \Rightarrow x \geq \frac{1}{2^m}$  and  $a_i = 0 \Rightarrow x \leq 0$ 
7 Then the s.stateShiftLeft stage implements:
8  $x = \text{LSig}(2^m x)$  // at that point  $a_i = 1 \Rightarrow x = 1$  and  $a_i = 0 \Rightarrow x = 0$ 
9 Then: the s.addToNumber stage updates the result accordingly – adding  $\frac{1}{4^i}$  or  $\frac{3}{4^i}$ , and so is the s.numberLeftToProcess.
10 */
11
12 Initialize:
13 [1] s.switch = 0
14 [2] s.switchTurnedOff = 0
15 [3] s.messageBitLength = 0
16 [4] s.numberLeftToProcess = 0
17 [5] s.numberInC4 = 0
18 [6] s.number_in_process = 0
19 [7] s.stateInit = 0
20 [8] s.stateReduce = 0
21 [9] s.stateShiftLeft = 0
22 [10] s.stateAddToNumber = 0
23 [11] s.maxDigitsToTheRight = 0
24 [12] s.digitsToTheLeft = 0
25 [13] s.digitsToTheRight = 0
26 [14] s.nextReduce = 0
27 [16] s.C41 = 0
28 [17] s.C43 = 0
29
30
31 RecurrentOp:
32 var switchWasOff = LSig(s.stateReduce + s.stateShiftLeft + s.stateAddToNumber + s.stateInit)
33 var switchTurnedOn = ReLU(s.switch - switchWasOff)
34 s.stateInit = switchTurnedOn
35 var prevSwitchVal = s.switch
36
37 // determine state of current pass
38 s.stateReduce = LSig(LSig(s.stateReduce - LSig(1 - s.digitsToTheRight)) + \lsig(s.stateInit - \lsig(s.messageBitLength -
39   ↪ s.maxDigitsToTheRight)) - (1 - s.switch))
40 s.stateShiftLeft = LSig(1 - s.stateReduce - s.stateInit - (1 - s.switch))
41 s.stateShiftLeft = LSig(s.stateShiftLeft - LSig(1 - s.digitsToTheLeft) - (1 - s.switch))
42 s.stateAddToNumber = LSig(LSig(s.maxDigitsToTheRight) - (s.stateReduce + s.stateShiftLeft + s.stateInit) - (1 - s.switch))
43
44 s.stateInit = LSig(LSig(s.messageBitLength - s.maxDigitsToTheRight) - (1 - s.stateInit) - (1 - s.switch))
45 // operation in init mode
46 var change = LSig(s.messageBitLength - s.digitsToTheRight)
47 s.digitsToTheRight.increase_if(s.stateInit, change)
48 s.maxDigitsToTheRight.increase_if(s.stateInit, 1)
49 s.nextReduce.set_if(s.stateInit, 1 / 4.0)

```

```

49 s_C41.set_if(s_stateInit, 1 / 4.0)
50 s_C43.set_if(s_stateInit, 3 / 4.0)
51 s_numberInC4.set_if(s_stateInit, 0)
52
53 // operation in the reduce state
54 s_digitsToTheRight.decrease_if(s_stateReduce, 1)
55 s_digitsToTheLeft.increase_if(s_stateReduce, 1)
56 var tmp = s_number_in_process
57 tmp.decrease_if(s_stateReduce, s_nextReduce)
58 s_number_in_process = LSig(tmp)
59 s_nextReduce.div2_if(s_stateReduce)
60
61 // operation in the shift left state
62 var tmp2 = s_number_in_process
63 tmp2.increase_if(s_stateShiftLeft, s_number_in_process)
64 s_number_in_process = LSig(tmp2)
65 s_digitsToTheLeft.decrease_if(s_stateShiftLeft, 1)
66 s_nextReduce.increase_if(s_stateShiftLeft, s_nextReduce)
67 s_digitsToTheRight.increase_if(s_stateShiftLeft, LSig(s_maxDigitsToTheRight - 1 - s_digitsToTheRight))
68
69 // operation in add to number state
70 // multiplying s_number_in_process by 2, this is sometimes required to make it  $\geq 1$  (all when the extracted digit is 1)
71 var tmp3 = s_number_in_process
72 tmp3.increase_if(s_stateAddToNumber, s_number_in_process)
73 s_number_in_process = LSig(tmp3)
74 s_numberInC4.increase_if(s_stateAddToNumber, LSig(s_C41 - s_number_in_process) + LSig(s_C43 - (1 - s_number_in_process)))
75 s_C41.div4_if(s_stateAddToNumber)
76 s_C43.div4_if(s_stateAddToNumber)
77 s_numberLeftToProcess.decrease_if(s_stateAddToNumber, LSig(0.5 - (1 - s_number_in_process)))
78 s_numberLeftToProcess = LSig(s_numberLeftToProcess)
79 s_numberLeftToProcess.increase_if(s_stateAddToNumber, s_numberLeftToProcess)
80 s_number_in_process.decrease_if(s_stateAddToNumber, s_number_in_process)
81 s_number_in_process.increase_if(s_stateAddToNumber, s_numberLeftToProcess)
82
83 s_maxDigitsToTheRight.decrease_if(s_stateAddToNumber, 1)
84 s_stateReduce.increase_if(s_stateAddToNumber, s_stateReduce)
85
86 s_switch = LSig(s_stateReduce + s_stateShiftLeft + s_stateAddToNumber + s_stateInit)
87 s_switchTurnedOff = LSig(prevSwitchVal - s_switch)

```

Listing 5: Implement Q2B Translation

```

1 Initialize:
2   [1] s_switchOn = 0
3   [2] s_switchTurnedOff = 0
4   [3] s_numberLeftToProcess = 0
5   [4] s_numberInBinary = 0
6   [5] s_stateAddToNumber = 0
7   [6] s_nextPosBinaryValue = 0
8
9 RecurrentOp:
10  var prevSwitchVal = s_stateAddToNumber
11  s_stateAddToNumber = \lsig(LSig(8 - s_numberLeftToProcess - 1) - ReLU(1 - s_switchOn))
12  s_switchOn = s_stateAddToNumber
13  s_switchTurnedOff = \relu(prevSwitchVal - s_switchOn)
14  var switchTurnedOn = \relu(s_switchOn - prevSwitchVal)
15
16  \\ init procedure, when switch turns on
17  s_nextPosBinaryValue.set_if(switchTurnedOn, 0.5)
18  s_numberInBinary.set_if(switchTurnedOn, 0)
19
20  \\ number translation procedure
21  var extractedDigit = LSig(4 * s_numberLeftToProcess - 2) // first digit is in {1,3} and we translate to {0,1}
22  s_numberInBinary.increase_if(s_stateAddToNumber, \lsig(s_nextPosBinaryValue - (1 - extractedDigit)))
23

```

```

24 s_numberLeftToProcess.decrease_if(s_stateAddToNumber, (1 + 2· extracted)/4.0)
25 // next 2 lines essentially multiply by 4
26 s_numberLeftToProcess = ChangelfState(s_stateAddToNumber, s_numberLeftToProcess) // multiply by 2
27 s_numberLeftToProcess = ChangelfState(s_stateAddToNumber, s_numberLeftToProcess) // multiply by 2
28 s_nextPosBinaryValue = div2_if_state(s_stateAddToNumber)

```

Listing 6: Implement Translation Of Final Result

```

1 Initialize:
2   [1] s_switchOn = 0
3   [2] s_switchTurnedOff = 0
4   [3] s_allFinished = 0
5   [4] s_numberLeftToProcess = 0
6   [5] s_allFinishedZeroOne = 0
7   [6] s_numberInBinary = 0
8   [7] s_stateAddToNumber = 0
9   [8] s_nextPosBinaryValue = 0
10  [9] s_lockResult = 0 // once we have the final result (in dimensions 5,6) we want it to stay no matter what happens in the network.
11
12 RecurrentOp:
13   var prevSwitchVal = s_stateAddToNumber
14   var allFinishedZeroOne = ReLU(4· s_allFinished - 2) // from RQ to {0,1}
15   var switchOnAndAllFinished = s_switchOn + allFinishedZeroOne - 1
16   s_stateAddToNumber = ReLU(LSig(LSig(8·s_numberLeftToProcess - 1) - (1 - switchOnAndAllFinished)) - s_lockResult)
17   s_switchTurnedOff = ReLU(prevSwitchVal - s_switchOn +
18     ReLU((1 - prevSwitchVal) + (s_switchOn - switchOnAndAllFinished) - 1))
19   s_allFinishedZeroOne.set_if(LSig(ReLU(s_switchTurnedOff + allFinishedZeroOne - 1) + s_lockResult), 1)
20   s_lockResult = \relu(s_lockResult + s.allFinishedZeroOne)
21   s_switchOn = s.stateAddToNumber
22   var shouldInit = ReLU(s_switchOn - prevSwitchVal)
23
24   \\ init procedure, when switch turns on
25   s_nextPosBinaryValue.set_if(shouldInit, 0.5)
26   s_numberInBinary.set_if(shouldInit, 0)
27
28   \\ number translation procedure
29   var extractedDigit = LSig(4· s_numberLeftToProcess - 2) // first digit is in {1,3} and we translate to {0,1}
30   s_numberInBinary.increase_if(s_stateAddToNumber, \lsig(s_nextPosBinaryValue - (1 - extracted)))
31   s_numberLeftToProcess.decrease_if(s_stateAddToNumber, (1 + 2· extracted)/4.0)
32 // next 2 lines essentially multiply by 4
33   s_numberLeftToProcess = ChangelfState(s_stateAddToNumber, s_numberLeftToProcess) // multiply by 2
34   s_numberLeftToProcess = ChangelfState(s_stateAddToNumber, s_numberLeftToProcess) // multiply by 2
35   s_nextPosBinaryValue = div2_if_state(s_stateAddToNumber)
36

```

Listing 7: Implement Synchronizer (F_S)

```

1 Initialize:
2   [1] s_stateReadyForNextAlgoIteration = 0 // signals the initiation of a new computation – starting from B2Q
3   [2] s_foundNotFinished = 0
4   [3] s_syncCountdown = 0
5   [4] s_stateSyncInProgress = 0
6   [5] s_stateCountdownOver = 1
7   [6] s_stateResetCountdown = 0
8   [7] s_syncNumOfCycles = 0 // external input, should be constant
9   [8] s_otherNodesNotFinished = 0 // external input
10  [9] s_curNodeFinished = 0 // external input
11
12 RecurrentOp:
13   s_stateReadyForNextAlgoIteration = 0; // updated during the pass
14   var cannotStart = ReLU(1 - s_syncNumOfCycles) // as long as s_syncNumOfCycles is not received, cannot start
15   var startedResetAndNotFinished = LSig( s_stateResetCountdown + LSig( s_syncRoundsCount - s_syncCountdown) - 1)
16   s_stateResetCountdown = ReLU(startedResetAndNotFinished - cannotStart) // do nothing until can start
17

```

```

18
19 var countdownGEzeroNotInReset = LSig(LSig( s.syncCountdown)−s.stateResetCountdown)
20 s.stateSyncInProgress = ReLU(countdownGEzeroNotInReset−cannotStart) // do nothing until can start
21 s.stateCountdownOver = ReLU $\left( (1 - \text{LSig}(s.syncCountdown)) - \text{cannotStart} \right)$  // ...
22
23 s.foundNotFinished = ReLU $\left( \text{LSig}(s.foundNotFinished + s.otherNodesNotFinished + (1 - s.curNodeFinished) - \right.$ 
    ↪  $s.stateResetCountdown) - \text{cannotStart} \right)$  // either already had not-finished indication in this cycle, or received indication
    ↪ of a non-finished node, or current node is not finished
24 s.syncCountdown.decrease_if(s.stateSyncInProgress, 1) // if during sync, countdown
25
26 s.stateReadyForNextAlgoIteration = LSig(s.stateCountdownOver+(1-s.foundNotFinished) − 1) // will be 0 while cannotStart
27 s.stateResetCountdown = s.stateReadyForNextAlgoIteration // reached 0 and all finished, should start reset
28 s.foundNotFinished = LSig(s.foundNotFinished−s.stateResetCountdown) // reset also resets foundNotFinished
29 s.syncCountdown.increase_if(s.stateResetCountdown, 1) // in reset mode continue to increase countdown

```

C Further Results

Theorem C.1. *There exists an mp-invariant graph embedding $F : \mathcal{G}_B \rightarrow \mathcal{B}$ such that for every R-GNN N there exists a disconnected graph G for which $N(G) \neq F(G)$.*

Proof. For all $(m, n) \in \mathbb{N}$ with $m, n \geq 3$, we define a graph $G_{m,n}$ to be the disjoint union of a cycle C_m^0 of length m in which all nodes have initial feature 0, and a cycle C_n^1 of length n in which all nodes have the initial feature 1.

We define a function $F : \mathcal{G}_B \rightarrow \mathcal{B}$ by

$$F(G_{m,n}) := \begin{cases} 1 & \text{if } m \text{ is even} \\ 0 & \text{if } m \text{ is odd} \end{cases}$$

for all $m, n \geq 3$ and $F(G) := 0$ if G is not isomorphic to some $G_{m,n}$ for $m, n \geq 3$. Clearly, F is computable and mp-invariant.

Suppose for contradiction that there is a graph-level R-GNN N computing F . We consider the computation of N on a graph $G_{m,n}$. After the computation stops, all vertices $v \in V(C_m^0)$ will have the same feature vector $\mathbf{x}_{m+n} \in \mathbb{Q}^k$, and all vertices $w \in V(C_n^1)$ will have the same feature vector $\mathbf{y}_{m+n} \in \mathbb{Q}^k$. Here k is a constant only depending on N . The vectors \mathbf{x}_{m+n} and \mathbf{y}_{m+n} may depend on the order $m+n$ of the input graph $G_{m,n}$, but not on m and n individually. To compute the final output, N passes the aggregated vector $\frac{m\mathbf{x}_{m+n}+n\mathbf{y}_{m+n}}{m+n}$ as input to an MLP which will compute the output $M(\frac{m\mathbf{x}+\mathbf{y}}{m+n}) = \text{rb}(F_N(G_{m,n}))$.

Every MLP with ReLu activations computes a piecewise linear function. This means that we can partition \mathbb{Q}^k into finitely many convex polytopes Q_1, \dots, Q_q , and on each Q_j the restriction of the function M computed by our MLP is linear (see, for example, (Grohe et al. 2025)). Within each Q_i , there is an affine and hence convex subset R_i where M is 1 (possibly, R_i is empty). Thus there are finitely many convex subsets $R_1, \dots, R_q \subseteq \mathbb{Q}^k$ such that for all $\mathbf{z} \in \mathbb{Q}^k$ we have

$$M(\mathbf{z}) = 1/2 \iff \mathbf{z} \in \bigcup_{\substack{i=1 \\ \vdash R}}^q R_i.$$

Let $\ell \in \mathbb{N}$ such that $\lfloor \frac{\ell-6}{2} \rfloor > q$. For $3 \leq m \leq \ell-3$, let $\mathbf{z}_m := \frac{m\mathbf{x}_\ell + (\ell-m)\mathbf{y}_\ell}{\ell}$. On input $G_{m,\ell-m}$, the output of N is $\text{rb}^{-1}(M(\mathbf{z}_m))$. As $F(G_{m,n}) = 1 \iff m \text{ is even}$, this means that $\mathbf{z}_m \in R$ for all even m and $\mathbf{z}_m \notin R$ for all odd m . Since R is the union of q sets R_i and there are more than q even m between 3 and $\ell-3$, there are an $i \in [q]$ and even $m_1 < m_2$ such that $\mathbf{z}_{m_1}, \mathbf{z}_{m_2} \in R_i$. However, \mathbf{z}_{m_1+1} is a convex combination of \mathbf{z}_{m_1} and \mathbf{z}_{m_2} , and thus $\mathbf{z}_{m_1+1} \in R_i \subseteq R$. It follows that $N(G_{m_1+1, \ell-m_1-1}) = \text{rb}^{-1}(M(\mathbf{z}_{m_1+1})) = 1$, while $F(G_{m_1+1, \ell-m_1-1}) = 0$. \square

Theorem 5.1. *Let $\mathcal{CG}_B \subset \mathcal{G}_B$ be the subset of connected graphs in \mathcal{G}_B , and let $F : \mathcal{CG}_B \rightarrow \mathcal{B}$ be computable. Then, F is mp-invariant if and only if there exists an R-GNN N such that $\forall G \in \mathcal{CG}_B N(G) = F(G)$. Furthermore, if F is computable in time $T(n)$ and space $S(n)$, then N uses time $O(T(n)) + \text{poly}(n)$ and space $O(S(n)) + \text{poly}(n)$.*

Proof. Define a feature transformation $F_T : \mathcal{CG}_B \rightarrow \mathcal{Z}_B$, $F_T(G)(v) := F(G)$. The connectivity of the input, together with Lemma A.11, mean that the (graph embedding) mp-invariance of F implies (feature transformation) mp-invariance of F_T . By the latter and Theorem 4.2 we have that there is an R-GNN N' such that $\forall G \in \mathcal{CG}_B \forall v \in V(G) N'(G, v) = F_T(G)(v)$. Let $N = (N', x \mapsto x)$ be a graph-embedding R-GNN where the final MLP computes the identity function, then we have that

$$\begin{aligned} \forall G \in \mathcal{CG}_B N(G) &= \\ \text{rb}^{-1}\left(\frac{1}{|G|} \sum_{v \in V(G)} \text{rb}(N'(G, v))\right) &= \\ \text{rb}^{-1}(\text{rb}(F_T(G)(v))) &= F_T(G)(v) = F(G) \end{aligned}$$

\square

Corollary 5.2. *Let $\mathcal{CG}_B \subset \mathcal{G}_B$ be the subset of connected graphs in \mathcal{G}_B , and let $F : \mathcal{CG}_B \rightarrow \mathcal{Z}_B$ be computable in time $T(n)$ and space $S(n)$. Then, there exists an R-GNN N with RNI, such that F is computable by N . Furthermore, N uses time $O(T(n)) + \text{poly}(n)$, space $O(S(n)) + \text{poly}(n)$, and $O(n \log n)$ random bits.*

Proof. We can view the computation of an R-GNN with RNI as a two-stage process: Given a graph $G \in \mathcal{G}_B$, $G = \{V(G), E(G), \mathcal{B}, Z(G)\}$, we first extend the initial feature of every node by a random number. This gives us a graph which we denote by \tilde{G} , $\tilde{G} = \{V(G), E(G), \mathcal{B}, \tilde{Z}(G)\}$, with the same structure as G but extended features. Then, we run a deterministic R-GNN on \tilde{G} . As the features of an R-GNN are rational numbers - finite precision, we restrict the length of the binary representation of the random numbers to $3 \log n$, that is, we choose a random bitstring of length $3 \log n$ for each node. With probability greater than $2/3$, each node will get a different random number assigned to it, in which case \tilde{G} is considered *individualized*.

Let $G \in \mathcal{CG}_B$ be a connected graph, and let $v \in V(G)$. For \tilde{G}, v we define

$$\tilde{F}(\tilde{G})(v) := \begin{cases} F(G)(v) & \tilde{G} \text{ is individualized} \\ \text{null-value} & \text{otherwise} \end{cases}$$

Note that for any two augmented graphs \tilde{G}, \tilde{H} that are connected and individualized, and two vertices $v \in V(\tilde{G}), v' \in V(\tilde{H})$, it holds that $\text{mpc}_{\tilde{G}}(v) = \text{mpc}_{\tilde{H}}(v') \Rightarrow$ there is an isomorphism $f : V(\tilde{G}) \rightarrow V(\tilde{H})$ such that $f(v) = f(v')$. Hence, for such $\tilde{G}, v, \tilde{H}, v'$ it holds that $\text{mpc}_{\tilde{G}}(v) = \text{mpc}_{\tilde{H}}(v') \Rightarrow \tilde{F}(\tilde{G})(v) = \tilde{F}(\tilde{H})(v')$. Hence, \tilde{F} is mp-invariant. Hence, by Theorem 4.2 there is an R-GNN N' that computes \tilde{F} . Hence, let N be an RNI R-GNN that uses N' as its deterministic R-GNN, then N computes F . \square

Theorem 5.3. *Let $F : \mathcal{G}_B \rightarrow \mathcal{Z}_B$ be a computable feature transformation. Then F is WL-invariant if and only if there is an R-GNN with global aggregation that computes F . Furthermore, if F is computable in time $T(n)$ and space $S(n)$, then the R-GNN uses time $O(T(n)) + \text{poly}(n)$ and space $O(S(n)) + \text{poly}(n)$.*

Proof. The proof imitates the proof of Theorem 4.2, reducing WL-invariant functions to R-GNNs with global aggregation, using adapted versions of the intermediate models MPC-GA; MP-LGA; and S-MP-GA, which we will denote by MPC-GA^w;MP-LGA^w;and S-MP-GA^w:

- MPC-GA^w differs from MPC-GA in that it receives the dag representing $\text{wl}_G(v)$ instead of $\text{mpc}_G(v)$
- MP-LGA^w differs from MP-LGA in that in each iteration it has 3 inputs rather than 2 - the third input being the multiset of values of $V(G) \setminus N_G(v)$.
- S-MP-GA^w differs from S-MP-GA in that in each iteration it has 5 inputs rather than 4 - the additional input being the sum of values of $V(G)$.

The reduction from MPC-GA^w to MP-LGA^w is a straightforward adaptation of the reduction from MPC-GA to MP-LGA. Reducing MP-LGA^w to S-MP-GA^w is similar to reducing MP-LGA to S-MP-GA: It is not difficult to see how to modify the 'receive_max' procedure (see Listing 1) to use a new input 'global_sum' and, in addition to the current update of the multiset of neighbors values, update a multiset of the values of $V(G) \setminus N_G(v)$, thus collecting the required information for emulating an MP-LGA^w iteration. Finally, reducing S-MP-GA^w to R-GNN with global sum-aggregation can be achieved by replicating the processing of the neighbors' sum:

- Having a sub-network F_{B2Q_s} , similar to F_{B2Q} , to translate the received global sum from RB to RQ.
- Having dimensions in F_M to accommodate the global sum input, thus being able to emulate a Turing machine that has the same inputs as an S-MP-GA^w.
- Having connections between F_{B2Q_s} and other sub-networks, similar to those that F_{B2Q} has.

□